Nick Bostrom

Superinteligencia

caminos, peligros, estrategias

The New York Times
Best Sellers

"Recomiendo encarecidamente este libro"
Bill Gates

TEELL

3ª edición

Superinteligencia

caminos, peligros, estrategias

Superinteligencia

Caminos, peligros, estrategias

NICK BOSTROM
Director, Future of Humanity Institute
Profesor, Facultad de Filosofía y Oxford Martin School,
Universidad de Oxford

Publicación en inglés por OXFORD University Press Título Original: SUPERINTELLIGENCE, Paths, Dangers, Strategies
Derechos de autor ©2014 by Nick Bostrom All rights reserved

Superintelligence was originally published in English in 2014. This translation is published by arrangement with Oxford University Press. Teell Editorial is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

Superinteligencia fue publicada originalmente en inglés en 2014. Esta traducción se publica con acuerdo con Oxford University Press. Teell Editorial solamente es responsable de la traducción del trabajo original y Oxford University Press no tendrá
ninguna responsabilidad de cualquier error, omisión, ambigüedad o descuido en dicha
traducción o por cualquier pérdida causada por la dependencia de la misma.

EDITADO POR TEELL EDITORIAL, S.L.
Primera edición en español: Teell Editorial, S.L. 2016
www.teelleditorial.com
Traductor e Introducción: Marcos Alonso

Diseño cubierta: www.uypdesign.com

Reservados todos los derechos. Queda rigurosamente prohibida, sin la autorización escrita de los titulares del “Copyright" bajo las sanciones establecidas en las leyes, la reproducción parcial o total de esta obra por cualquier medio o procedimiento, incluida la reprografía y el tratamiento informático, así como la distribución de ejemplares mediante alquiler o préstamo público.

e ISBN: 978-84-16511-56-3
Depósito Legal: Z 70-2016
Impreso en España por Talleres Editoriales Cometa, S.A.

INTRODUCCIÓN

A

nte la publicación en castellano de la trascendental obra de Nick Bostrom, *Superinteligencia,* se hace necesario recordar las palabras del gran filósofo español José Ortega y Gasset: “Uno de los temas que en los próximos años se va a debatir con mayor brío es el del sentido, ventajas, daños y límites de la técnica" Estas proféticas palabras de principios del siglo XX no han hecho más que cumplirse, siendo el presente libro un verdadero hito en la reflexión sobre la técnica o tecnología. Justo después de las citadas palabras de *Meditación de la técnica,* Ortega expresaba que “la misión del escritor es prever con holgada anticipación lo que va a ser problema, años más tarde, para sus lectores y proporcionarles a tiempo, es decir, antes de que el debate surja, ideas claras sobre la cuestión" Eso es precisamente lo que Bostrom ha buscado con *Superinteligencia:* presentar de manera clara y accesible un tema complejo y decisivo como es el problema de la inteligencia artificial de nivel sobrehumano.

La preocupación por las máquinas inteligentes no es algo completamente nuevo. Desde la segunda mitad del siglo pasado se ha convertido en una parte central de nuestro imaginario colectivo, con numerosas novelas y películas dedicadas al tema. En *Superinteligencia* veremos cómo se discuten, desde una base científica sólida, temas como la posibilidad de un gobierno mundial automatizado (recordando a clásicos como *Un mundo feliz,* de Aldous Huxley o *1984,* de George Orwell); la plausible catástrofe para la humanidad derivada del mal funcionamiento de las máquinas (un tema recurrente en el cine, como atestiguan *2001 Odisea en el espacio,* de Stanley Ku- brick, *Terminator,* de James Cameron o *Matrix,* de los hermanos Wachowski); o las posibilidades y limitaciones de IAs con valores humanos (reflejadas en obras como *Yo, robot,* de Isaac Asimov o en la reciente *Her,* de Spike Jonze).

Las creaciones artísticas anticipan el mundo que está por venir, aunque en el caso de la IA los avances científico-técnicos suceden de una manera tan vertiginosa que la imaginación humana tiene dificultades para mantener el ritmo. El gran mérito de *Superinteligencia* es precisamente abordar de manera seria y rigurosa problemas que hasta hace no mucho parecían meras ocurrencias de ciencia-ficción. El libro de Bos- trom es el mayor intento habido hasta la fecha de afrontar estos problemas desde una perspectiva estrictamente filosófica y científica, poniendo en juego un conocimiento realmente enciclopédico sobre la materia; un esfuerzo que sólo un pensador de la talla de Bostrom podía llevar a cabo.

En este sentido, es importante hacer notar que Bostrom tampoco juega a ser adivino, como sí hacen algunos gurús de la tecnología actuales. Bostrom no categoriza sobre la llegada de la IA ni sobre sus consecuencias. Sin embargo, la fuerza de sus argumentos nos sobresalta y nos despierta de nuestro letargo, haciéndonos ver el inmenso peligro que traería consigo la llegada de una superinteligencia artificial. Esta toma de conciencia ni siquiera requiere por nuestra parte una gran fe en el progreso tecnológico; de hecho, el propio Bostrom se muestra escéptico respecto de las predicciones más optimistas sobre IA. Basta con atisbar el más que probable desastre al que abocaría un mal desarrollo de la IA para que nos preocupemos muy seriamente por estas tecnologías, confiemos plenamente en su desarrollo exitoso o no.

A pesar de lo que muchas veces se ha pensado —principalmente por su inclusión en la corriente del transhumanismo—, Nick Bostrom no es un tecnólatra. Tampoco un tecnófobo, aunque alguna lectura apresurada de *Superinteligencia* podría sacar esta conclusión. Bostrom es simplemente un pensador. Y como tal, su misión es reflexionar de manera radical sobre los problemas prominentes de su época. Su acierto consiste en haber identificado a la tecnología moderna como problema fundamental, y más concretamente a la fe ciega y el optimismo exacerbado que muchas personas profesan en la actualidad por la tecnología.

La obra lleva por subtítulo *Caminos, peligros, estrategias,* pero de estos términos acaba emergiendo uno por encima de los demás: el peligro. Necesitamos comprender los caminos que llevan a la superinteligencia para plantear las mejores estrategias frente a ella; pero la cuestión central en torno a la que giran todas las demás es el peligro que la superinteligencia trae consigo, la amenaza que proyecta sobre la humanidad y que deberíamos tener muy en cuenta. Este tono de advertencia no desemboca, en el caso de Bostrom, en un miedo paralizante que demonice todo avance en IA; únicamente nos hace comprender la enorme responsabilidad que tenemos y la imprudencia

que sería desentenderse de estos problemas o confiar ingenuamente en su desarrollo benéfico.

A continuación sobrevolaremos brevemente el libro para dar una pequeña guía de lectura así como para aprovechar para comentar algunos puntos y términos especialmente problemáticos.

Superinteligencia se divide básicamente en dos bloques: un primer grupo de capítulos dedicado a presentar el problema de la inteligencia artificial, atendiendo a su historia, sus posibilidades presentes y su previsible desarrollo en los próximos años; y un segundo bloque algo más extenso dedicado a pensar qué podremos hacer con la superinteligencia, los peligros y problemas que probablemente acarree, y las estrategias y soluciones a nuestro alcance.

El primer bloque se extiende aproximadamente hasta el capítulo 6. Tras un breve recorrido en el primer capítulo a la historia de la IA y una somera exposición del estado de la cuestión, Bostrom pasa, ya en el segundo capítulo, a exponer los posibles caminos que podrían llevar hasta la superinteligencia. Este capítulo es crucial en tanto que introduce algunos términos omnipresentes a lo largo del libro, principalmente los de IA (inteligencia artificial) y ECC (emulación de cerebro completo). Sobre este punto conviene advertir del uso que Bostrom hace de inteligencia artificial en un sentido amplio para referirse a cualquier forma no humana de inteligencia, y un uso más restringido de inteligencia artificial para referirse a los tipos de IA cuyo origen no está directamente relacionado con el cerebro humano, frente a los que sí tienen dicho órgano como base (como la emulación de cerebro completo).

En el capítulo 3, Bostrom considera brevemente los distintos tipos de superinte- ligencia que podrían surgir, algo que retomará en más profundidad en el capítulo 6, cuando hable de los diferentes superpoderes cognitivos que una superinteligencia presumiblemente tendría a su disposición. Bostrom introduce aquí por primera el decisivo concepto de Unidad, un ente único que concentraría todo el poder (político, tecnológico, económico, etc.) en sí mismo. Esta idea estará presente a lo largo de toda la obra y aparecerá repetidamente en numerosos párrafos.

Los capítulos 4 y 5 intentan atisbar la manera concreta en que la superinteligencia surgirá. Bostrom insiste en la relevancia del momento en que tenga lugar la explosión de inteligencia y en la importancia de su velocidad de despegue: cuanto antes suceda menos preparados estaremos. Bostrom propondrá una fórmula para calcular esta velocidad basada en la potencia de optimización y su contrapartida la resistencia al progreso (una construcción que traduce *recalcitrance,* término inexistente en castellano y cuyos sinónimos directos son inadecuados). La clave de estas discusiones y lo que más

interesa al autor es la posibilidad de que la superinteligencia adquiera una ventaja estratégica decisiva, otro concepto que sobrevolará toda la obra.

Si bien, como ya hemos dicho, *Superinteligencia* es un libro unitario y muy bien trabado, el capítulo 7 marca la entrada en lo que podríamos considerar la segunda parte del libro. Este capítulo aborda por primera vez el problema de las relaciones entre inteligencia y motivación, un marco conceptual que servirá de base al resto de la obra. Un primer gran peligro en este sentido es el de antropomorfizar la IA, asumiendo que cuanto más inteligente sea la IA, más humana será y sus motivaciones y objetivos más se parecerán a los humanos. Esto podría no ser así, y, de hecho, si no nos esforzamos específicamente por conseguirlo, lo más probable es que la superinteligencia que surja sea profundamente inhumana.

El octavo capítulo se pregunta, de manera directa y sin rodeos, si estamos abocados al desastre. Este tono de preocupación gobernará la segunda parte del libro, en la que Bostrom tratará de hacerse cargo de las posibles amenazas de la superinteligencia, proponiendo, a su vez, las que considera mejores estrategias para afrontarlas, principalmente la prevención. Como hemos dicho, esta prudencia no debe confundirse con una tendencia al catastrofismo por parte del autor; la lectura de *Superinteligencia* no permite sacar esa conclusión. Lo que encontramos es una fría y objetiva mirada a la realidad, en la que se descubren unos riesgos que deben ser tenidos en cuenta. En este capítulo 8, Bostrom presenta diversos modos concretos en que la IA podría fallar, como la suplantación perversa (una IA que interpretara defectuosamente nuestras órdenes, suplantándolas por otras con efectos perjudiciales) o el crimen mental (la posibilidad de que las IAs o las emulaciones tengan un estatus moral que podría ser violado).

Los capítulos 9 y 10 están dedicados a examinar una primera salida a los problemas de la superinteligencia: los métodos de control. Éstos se dividen, por un lado, en métodos de control de la capacidad, que buscan impedir que la superinteligencia tenga un poder efectivo total sobre el mundo; y, por otro lado, en métodos de selección de la motivación, que buscan elegir los objetivos que la superinteligencia llegaría a tener. Bostrom demuestra que ambos métodos tiene carencias. En el caso del control de capacidad el problema está en la tensión existente entre minar las capacidades de la IA lo suficiente para que no tenga un poder absoluto, sin restarle tanta capacidad que ya no pueda ser considerada superinteligente. En el capítulo 10, Bostrom comparará diversos tipos de IA en función de su idoneidad para el control. La selección de la motivación es un problema más complejo, que por ello se trata más detalladamente en los últimos capítulos, donde encontramos la verdadera clave de la propuesta de

Bostrom.

De este modo, la reflexión sobre los métodos de selección de la motivación conducen a Bostrom, en los capítulos 12 y 13, a reflexionar a fondo sobre la posibilidad de crear una IA con valores. Para Bostrom, ésta es la única salida que, en caso de ser posible, supondría una verdadera solución al problema de la IA superinteligente. Por eso el filósofo sueco concentra sus mejores esfuerzos en esta parte de la obra, que entiende como decisiva. El lector juzgará éxito de este intento, pero no cabe duda de que Bostrom es un gran conocedor tanto de las nuevas tecnologías como de filosofía moral. Sería imposible dar cuenta aquí del contenido de estos capítulos, pero sí merece especial atención la propuesta de la VCE (voluntad coherente extrapolada), una forma de normatividad indirecta que busca aprovechar la propia capacidad de la superinteligencia para llevar a cumplimiento de manera más perfecta los valores humanos comúnmente aceptados.

El último tema que aborda Bostrom es el del escenario estratégico que probablemente tenga lugar una vez que la superinteligencia se haga efectiva. Esto es estudiado en el capítulo 11, donde se empieza exponiendo las presumibles consecuencias sociales, políticas y económicas de la aparición de la superinteligencia, para luego comparar la plausibilidad de un escenario multipolar (con varias superinteligencias coexistiendo en igualdad) frente a un escenario de Unidad (en el que todos los poderes se fusionen en uno solo). Esta temática reaparece en el capítulo 14, donde el autor valora factores concretos que posiblemente configuren el surgimiento o establecimiento de la superinteligencia (como el desarrollo tecnológico diferencial y los acoplamientos tecnológicos) al tiempo que brinda soluciones concretas a esas situaciones, soluciones que retoman muchas de las discusiones en torno al problema de introducción de valores en la IA. Por último, el capítulo 15 sirve para recapitular las cuestiones fundamentales antes expuestas y hacer un último llamamiento a la concienciación sobre los problemas de la superinteligencia y un ruego a que le dediquemos nuestros mejores esfuerzos intelectuales en el futuro próximo.

Como expresa el autor en el prefacio, la pretensión de escribir un libro accesible sólo se ha conseguido parcialmente. Lo mismo puede decirse de la traducción. La concisión y claridad de lenguaje que Bostrom exhibe y que hemos intentado transmitir no impide que haya abundantes pasajes técnicos e intrincados, así como algunos términos de difícil traducción. En algunos casos, los menos, se ha mantenido el inglés original, principalmente cuando nos encontrábamos con términos habituales en la bibliografía de este tipo (un ejemplo sería wetware, un término utilizado en referencia a estructuras “húmedas”,

de origen orgánico, en oposición a estructuras inorgánicas como el software o el hardware). En otros casos, como el mencionado de Unidad (traducción del original Singleton), se ha preferido crear un término que diera a entender la idea original del autor. Las notas e índice situados al final de la obra sirven también de gran ayuda, y el lector no debería dudar en consultarlos.

En cualquier caso y como decíamos al principio, *Superinteligencia* es una gran obra no sólo por la importancia crucial de su tema y por la excelencia de su ejecución, sino también por su vocación de llegar al gran público con un lenguaje cercano y una estructura fácilmente inteligible. *Superinteligencia* es una obra destinada a ser leída, discutida y tenida en cuenta. No es una exageración decir que, hasta cierto punto, el futuro de la humanidad depende de ello.

MARCOS ALONSO FERNÁNDEZ

La inacabada fábula de los gorriones

Era la temporada de construcción de nidos, pero después de días de largo y duro trabajo, los gorriones se sentaron a la luz del atardecer, relajándose y trinando. “Somos todos tan pequeños y débiles. ¡Imaginad lo fácil que sería la vida si tuviéramos una lechuza que nos ayudara a construir nuestros nidos!”

“¡Sí!”, dijo otro. “Y podría ayudarnos a cuidar de nuestros ancianos e hijos”. “Podría darnos consejo y vigilar a los gatos que merodean cerca de aquí”, añadió un tercero.

Entonces Pastus, el pájaro anciano y sabio, habló: “Enviemos exploradores en todas las direcciones e intentemos encontrar una lechuza joven abandonada en alguna parte, o quizás un huevo. Una cría de cuervo también serviría, o una comadreja muy joven. Podría ser lo mejor que jamás nos hubiera sucedido, al menos desde la apertura del «pabellón de grano ilimitado» del patio trasero”.

La multitud estaba entusiasmada, y por todas partes los gorriones comenzaron a trinar con toda la fuerza de sus pulmones.

Sólo Scronkfinkle, un gorrión con un solo ojo y de temperamento inquisitivo, no estaba convencido de la sensatez de la empresa. Y dijo: “Eso seguramente sea nuestra perdición. ¿No deberíamos pensar antes en la labor de domesticar y controlar a la lechuza, antes de meter una criatura así entre nosotros?”

Pastus contestó: “Domesticar una lechuza parece una cosa muy difícil. Ya será bastante complicado encontrar un huevo de lechuza. Así que empecemos con eso. Cuando logremos criar a una lechuza, nos pondremos a pensar en emprender esa otra tarea”.

“¡El plan tiene un defecto!”, pio Scronkfinkle; pero sus protestas fueron en vano, pues la multitud ya había alzado el vuelo para empezar a llevar a cabo el plan ideado por Pastus.

Sólo dos o tres gorriones se quedaron rezagados. Juntos empezaron a intentar adivinar cómo se podría domesticar y controlar a una lechuza. Pronto se dieron cuenta de que Pastus tenía razón: era un reto extremadamente difícil, sobre todo por carecer de una lechuza real con la que practicar. No obstante, continuaron haciéndolo lo mejor que pudieron, con el temor constante a que la multitud regresara con un huevo de lechuza antes de encontrar una solución al problema de cómo controlar esa criatura.

No se sabe cómo termina la historia, pero el autor del presente libro lo dedica a Scronfinkle y sus seguidores.

PREFACIO

D

entro de tu cráneo está la cosa que lee. Esa cosa, el cerebro humano, tiene algunas capacidades de las que carecen los cerebros de otros animales. Es debido a estas capacidades distintivas que nuestra especie ocupa una posición dominante sobre el planeta. Otros animales tienen músculos más fuertes o garras más afiladas, pero nosotros tenemos cerebros más inteligentes. Nuestra modesta ventaja en inteligencia general nos ha llevado a desarrollar lenguaje, tecnología y una compleja organización social. La ventaja ha ido aumentando con el paso del tiempo, ya que cada generación se ha basado en los logros de sus predecesores.

Si algún día llegáramos a construir cerebros artificiales que superaran en inteligencia general a los cerebros humanos, entonces esa nueva superinteligencia podría llegar a ser muy poderosa. Y, de igual manera que el destino de los gorilas depende ahora más de los humanos que de ellos mismos, también el destino de nuestra especie pasaría a depender de las acciones de la superinteligencia artificial.

No obstante, contamos con una ventaja: somos nosotros los que construiremos todo. En principio, podríamos construir un tipo de superinteligencia que protegiera los valores humanos. Sin duda, tendríamos poderosas razones para ello. En la práctica, el problema del control —el problema de cómo controlar qué haría esa superinteligencia— parece bastante complicado. También parece evidente que sólo tendremos una única oportunidad. Si alguna vez llegara a existir una superinteligencia poco amistosa, nos impediría sustituirla o cambiar sus preferencias. Nuestro destino estaría sellado.

En este libro, mi objetivo es entender el reto ofrecido por la perspectiva de la superinteligencia, y cuál sería nuestra mejor respuesta. Se trata, muy probablemente, del reto más importante y sobrecogedor al que la humanidad se haya enfrentado nunca. Y —tanto si tenemos éxito como si fracasamos—, probablemente sea el último reto que tengamos que afrontar.

No se discute en este libro que nos encontramos en los umbrales de un gran avance en el campo de la inteligencia artificial, ni que podamos predecir con ningún grado de exactitud cuándo podría tener lugar tal desarrollo. Hay cierta probabilidad de que ocurra en algún momento de este siglo, pero no lo sabemos con seguridad. Los dos primeros capítulos tratan sobre posibles caminos y dicen algo sobre su desarrollo temporal. No obstante, la mayor parte del libro trata sobre lo que ocurrirá después. Estudiamos la cinética de una explosión de inteligencia, las formas y poderes de la superinteligencia, y las decisiones estratégicas disponibles para un ser superinteligen- te que obtuviera una ventaja decisiva. Después centraremos nuestra atención en el problema del control y nos preguntaremos qué podríamos hacer para configurar las condiciones iniciales, con el objetivo de lograr un resultado beneficioso con el que pudiéramos sobrevivir. Hacia el final del libro, nos acercaremos y contemplaremos el panorama general que emerge de nuestras investigaciones. Se ofrecerán algunas recomendaciones sobre lo que debemos hacer en la actualidad a fin de aumentar nuestras posibilidades de evitar una catástrofe existencial posterior.

Éste no ha sido un libro fácil de escribir: espero que el camino que se ha abierto permita a otros investigadores alcanzar la nueva frontera más suave y fácilmente, para que puedan llegar allí frescos y preparados para sumarse al trabajo de extender nuestra comprensión. (Y si el camino que se ha abierto es un poco accidentado y sinuoso, ¡espero que los críticos no subestimen lo hostil que el terreno era *ex ante!)*.

Éste no ha sido un libro fácil de escribir: he tratado que fuera un libro sencillo de leer, pero no creo haber tenido excesivo éxito. Al escribir, imaginaba al posible lector como una antigua parte de mí mismo, e intenté redactar el tipo de libro que habría disfrutado leyendo. Este tipo de persona podría ser una parte muy reducida de la población. Sin embargo, creo que el contenido debe ser asequible a muchas personas, si reflexionan un poco sobre ello y resisten la tentación de malinterpretar automáticamente cada nueva idea reduciéndola al estereotipo más más similar disponible en su despensa cultural. A los lectores a los que cueste entender las secciones más técnicas no debería desanimarles algún que otro fragmento con matemáticas o vocabulario especializado, ya que siempre es posible deducir el tema principal de las explicaciones que lo acompañan. (A la inversa, aquellos lectores que deseen tener algo más allá de lo básico, podrán encontrar bastante material entre las notas finales[1]).

Muchos de los puntos señalados en este libro probablemente sean erróneos.[2] También es probable que haya consideraciones de importancia crítica que yo no haya tenido en cuenta, lo cual

invalidaría todas o algunas de mis conclusiones. He indicado en cierta medida los matices y grados de incertidumbre a lo largo de todo el texto, señalándolos con las feas marcas de “posiblemente”, “podría”, “puede”, “bien podría”, “parece”, “probablemente”, “muy probablemente” y “casi con total seguridad”. Cada uno de estos modificadores se ha colocado en su lugar correspondiente de manera cuidadosa y deliberada. Sin embargo, estas aplicaciones tópicas de modestia epistemológica no son suficientes; deben ir acompañadas por la admisión sistemática de incertidumbre y falibilidad. Esto no es falsa modestia, porque aunque creo que mi libro es probablemente erróneo y engañoso, pienso que los puntos de vista alternativos que se han presentado en la literatura científica son sustancialmente peores — incluyendo el enfoque por defecto, o “hipótesis nula”, según la cual podemos de momento ignorar de forma segura o razonable la perspectiva de la superinteligencia.

AGRADECIMIENTOS

L

a membrana que ha recubierto el proceso de redacción de este libro ha sido bastante permeable. Muchos conceptos e ideas generados mientras trabajaba en él han ido filtrándose y se han convertido en parte de una conversación más amplia; y, por supuesto, se han incluido numerosas ideas procedentes del exterior que aparecieron mientras escribía el libro. He intentado ser bastante diligente con las citas, pero hay demasiadas influencias como para poder documentarlas de manera exhaustiva.

Por los prolongados debates que han ayudado a aclarar mi pensamiento estoy agradecido a una larga serie de personas; entre ellas, a Ross Andersen, Stuart Ar- mstrong, Owen Cotton-Barratt, Nick Beckstead, David Chalmers, Paul Christiano, Milan Cirkovic, Daniel Dennett, David Deutsch, Daniel Dewey, Eric Drexler, Peter Eckersley, Amnon Eden, Owain Evans, Benja Fallenstein, Alex Flint, Carl Frey, Ian Goldin, Katja Grace, J. Storrs Hall, Robin Hanson, Demis Hassabis, James Hughes, Marcus Hutter, Garry Kasparov, Marcin Kulczycki, Shane Legg, Moshe Looks, Wi- llam MacAskill, Eric Mandelbaum, James Martin, Lillian Martin, Roko Mijic, Vincent Mueller, Elon Musk, Seán Ó hÉigeartaigh, Toby Ord, Dennis Pamlin, Derek Parfit, David Pearce, Huw Price, Martin Rees, Bill Roscoe, Stuart Russell, Anna Salamon, Lou Salkind, Anders Sandberg, Julian Savulescu, Jurgen Schmidhuber, Nicholas Shac- kel, Murray Shanahan, Noel Sharkey, Carl Shulman, Peter Singer, Dan Stoicescu, Jaan Tallinn, Alexander Tamas, Max Tegmark, Roman Yampolskiy y Eliezer Yudkowsky.

Por sus comentarios especialmente detallados, estoy agradecido a Milan Cirkovic, Daniel Dewey, Owain Evans, Nick Hay, Keith Mansfield, Luke Muehlhauser, Toby Ord, Jess Riedel, Anders Sandberg, Murray Shanahan y Carl Shulman. Por sus consejos o su ayuda en la investigación de distintas partes quiero dar las gracias a Stuart Armstrong, Daniel Dewey, Eric Drexler, Alexandre Erler, Rebecca Roache y Anders Sandberg.

Por su ayuda en la preparación del manuscrito, estoy agradecido a Caleb Bell, Malo Bourgon, Robin Brandt, Lance Bush, Cathy Douglass, Alexandre Erler, Kristian Ronn, Susan Rogers, Andrew Snyder-Beattie, Cecilia Tilli y Alex Vermeer. Deseo dar las gracias especialmente a mi editor, Keith Mansfield, por todos los ánimos que me ha dado a lo largo del proyecto.

Mis disculpas a todo aquel que debería haber sido recordado aquí.

Finalmente, mi agradecimiento más afectuoso a los financiadores, amigos y familia: sin vuestro apoyo, este trabajo no habría tenido lugar.

INDICE DE CONTENIDOS

Listas de figuras, tablas y cuadros xix

1. Desarrollos del pasado y capacidades del presente 1
Modos de crecimiento y la historia a gran escala 1
Grandes expectativas 3
Épocas de esperanza y desesperación 5
Estado de la técnica 11
Opiniones sobre el futuro de la inteligencia artificial 18

2. Caminos hacia la superinteligencia 22
Inteligencia artificial 23
Emulación de cerebro completo 30
Cognición biológica 36
Interfaces cerebro-ordenador 44
Redes y organizaciones 48
Resumen 50

3. Formas de superinteligencia 52
Superinteligencia de velocidad 52
Superinteligencia colectiva 53
Superinteligencia de calidad 56
Alcance directo e indirecto 57
Fuentes de ventaja para la inteligencia digital 59

4. La cinética de una explosión de inteligencia 62
Sincronización y velocidad del despegue 62
Resistencia al progreso 66
Caminos que no implican la inteligencia artificial 66
Caminos a través de la emulación y la IA 68
Potencia de optimización y explosividad 74

5. Ventaja estratégica decisiva 78

¿Conseguirá el proyecto adelantado una ventaja estratégica decisiva? 79

¿Qué magnitud tendrá el proyecto exitoso? 83

Monitorización 84

Colaboración internacional 86

De una ventaja estratégica decisiva a la Unidad 87

6. Superpoderes cognitivos 91

Funcionalidades y superpoderes 91

Un escenario de toma de poder por parte de la IA 95

Poder sobre la naturaleza y sobre los agentes 99

7. La voluntad superinteligente 105

La relación entre inteligencia y motivación 105

Convergencia instrumental 109

Auto-conservación 109

Integridad del contenido de los objetivos 109

Mejora cognitiva 111

Perfección tecnológica 112

Adquisición de recursos 113

8. ¿Es el apocalipsis el resultado inevitable? 115

¿La catástrofe existencial como resultado predeterminado de una explosión de inteligencia? 115

El giro traicionero 116

Modos de fallo malignos 119

Suplantación perversa 120

Profusión infraestructural 122

Crimen mental 125

9. El problema del control 127

Problemas de agencia doble 127

Métodos de control de la capacidad 129

Métodos de encajamiento 129

Métodos de incentivos 131

Atrofia 135

Cables trampa 136

Métodos de selección de la motivación 138

Especificación directa 139

Domesticidad 140

Normatividad indirecta 141

Aumentación 142

Sinopsis 143

10. Oráculos, genios, soberanos, herramientas 145

Oráculos 145
Genios y soberanos 148
IAs-herramienta 150
Comparación 155

11. Escenarios multipolares 159

De caballos y hombres 160
Los salarios y el desempleo 160
El capital y el bienestar 161
El principio malthusiano desde una perspectiva histórica 163
Crecimiento demográfico e inversión 164
La vida en una economía algorítmica 166
Esclavitud voluntaria, muerte ocasional 167
¿El trabajo máximamente eficiente sería divertido? 170
¿Subcontratados inconscientes? 172
La evolución no es necesariamente hacia adelante 174
¿Formación post-transición de una Unidad? 177
Una segunda transición 177
Superorganismos y economías de escala 178
Unificación por tratado 180

12. Adquiriendo valores 185

El problema de la introducción de valores 185
Selección evolutiva 187
Aprendizaje por refuerzo 188
Acumulación de valores por asociación 189
Andamiaje motivacional 191
Aprendizaje de valores 192
Modulación de emulaciones 201
Diseño institucional 202
Sinopsis 207

13. Eligiendo los criterios para elegir 209

La necesidad de normatividad indirecta 209
Voluntad coherente extrapolada 211
Algunas explicaciones 212
Razones para la VCE 213
Otras observaciones 216
Modelos de moralidad 217
Haz lo que quiero que hagas 220
Lista de componentes 221

Contenido de los objetivos 222
Teoría de la decisión 223
Epistemología 224
Ratificación 225
Acercándonos lo suficiente 227

14. El panorama estratégico 229
Estrategia científica y tecnológica 229
Desarrollo tecnológico diferencial 230
Preferencia en el orden de llegada 231
El ritmo de cambio y la mejora cognitiva 234
Acoplamientos tecnológicos 237
Anticipándose a las consecuencias 239
Caminos y posibilitadores 241
Efectos de los avances de hardware 241
¿Debería promoverse la investigación sobre la emulación de cerebro completo? 242
La perspectiva de la persona afectada favorece la aceleración 246
Colaboración 247
La dinámica de carrera y sus peligros 247
Sobre los beneficios de la colaboración 250
Trabajando juntos 254

15. La hora de la verdad 256
Filosofía con fecha límite 256
¿Qué debemos hacer? 257
Buscar la luz estratégica 258
Desarrollar una buena capacidad 258
Medidas particulares 259
Que los mejores en naturaleza humana por favor se pongan en pie 260

Notas 263
Bibliografía 307
Índice 327

LISTAS DE FIGURAS, TABLAS Y CUADROS

Lista de Figuras

1. Historia a largo plazo del producto interior bruto mundial (GDP). 3
2. Impacto global a largo plazo de la inteligencia artificial de nivel humano. 21
3. Rendimiento de las supercomputadoras. 27
4. Reconstrucción 3D neuroanatómica a partir de imágenes de un microscopio electrónico. 31
5. Hoja de ruta de la emulación de cerebro completo. 34
6. Las caras compuestas como una metáfora de corrección ortográfica para genomas. 41
7. Configuración del despegue. 63
8. ¿Una escala menos antropomórfica? 70
9. Un sencillo modelo de una explosión de inteligencia. 77
10. Fases de un escenario de toma de control por parte de una IA. 96
11. Esquema de posibles trayectorias encaminadas a una hipotética Unidad-sabia. 101
12. Resultados de antropomorfizar la motivación ajena. 106
13. ¿Inteligencia artificial, o antes simulación integral del cerebro? 244
14. Niveles de riesgo en las carreras tecnológicas de la IA. 248

Lista de tablas

1. IAs dedicadas a juegos. 12
2. ¿Cuándo conseguiremos una inteligencia artificial de nivel humano? 20
3. ¿Cuánto se tardará desde el nivel humano hasta la superinteligencia? 21
4. Capacidades necesarias para la emulación de cerebro completo. 32
5. Aumentos máximos del CI al seleccionar de entre un conjunto de embriones. 37
6. Posibles impactos de la selección genética en diferentes escenarios. 40

7. Algunas competiciones tecnológicas estratégicamente significativas. 81
8. Superpoderes: algunas tareas estratégicamente relevantes y su conjunto
de habilidades correspondientes. 94
9. Diversos tipos de cables de trampa. 137
10. Métodos de control. 143

11. Características de las diferentes castas del sistema. 156
12. Resumen de las técnicas de introducción de valores. 207
13. Lista de componentes. 222

Lista de cuadros

1. Un agente bayesiano óptimo. 10
2. El Flash Crash de 2010. 17
3. ¿Qué se necesitaría para recapitular la evolución? 25
4. Sobre la cinética de una explosión de inteligencia. 75
5. Carreras tecnológicas: algunos ejemplos históricos. 80
6. El escenario del ADN encargado por correo. 98
7. ¿Qué magnitud tienen los recursos cósmicos? 101
8. Captura antrópica. 134
9. Soluciones extrañas de búsquedas a ciegas. 154
10. Formalización de aprendizaje de valores. 194
11. Una IA que quiere ser amigable. 197
12. Dos ideas recientes (sin perfilar). 198
13. Una carrera arriesgada en picado. 248

CAPÍTULO 1

Desarrollos del pasado y capacidades actuales

Empezaremos mirando atrás. La historia, a gran escala, parece exhibir una secuencia de modos de crecimiento distintos, cada uno mucho más rápido que el anterior. Este patrón ha llevado a sugerir que otro modo de crecimiento (aún más rápido) podría ser posible. Sin embargo, no daremos mucho peso a esta observación —éste no es un libro sobre la "aceleración tecnológica" o el "crecimiento exponencial" o las diversas nociones a veces reunidas bajo el título de "la singularidad". A continuación, se revisará la historia de la inteligencia artificial. Después examinaremos las capacidades actuales de dicho campo. Por último, echaremos un vistazo a algunas recientes encuestas de opinión de expertos y reconoceremos nuestra ignorancia acerca de la línea temporal de los futuros avances.

Modos de crecimiento y la historia a gran escala

Hace apenas unos pocos millones de años nuestros antepasados todavía estaban colgando de las ramas de la selva africana. En una escala temporal geológica o incluso evolutiva, el surgimiento del *Homo sapiens* a partir de nuestro último ancestro común con los grandes simios tuvo lugar rápidamente. Desarrollamos la postura erguida, los pulgares oponibles y —de manera crucial— algunos cambios relativamente menores en el tamaño del cerebro y en la organización neurológica que conllevaron un gran salto en la capacidad cognitiva. Como consecuencia, los seres humanos pueden pensar de manera abstracta, comunicar pensamientos complejos y acumular culturalmente información a lo largo de generaciones mucho mejor que cualquier otra especie sobre el planeta.

Estas capacidades permitieron a los seres humanos desarrollar tecnologías productivas cada vez más eficientes, haciendo posible que nuestros antepasados emigraran lejos de la selva y la sabana. Especialmente después de la adopción de la agricultura, la densidad de población aumentó junto con el tamaño total de la población humana. Que hubiera más personas conllevó que hubiera más ideas; las mayores densidades de población permitieron que las ideas se

propagaran con mayor facilidad y que algunos individuos pudieran dedicarse al desarrollo de habilidades especializadas. Estos acontecimientos aumentaron el *ritmo de crecimiento* de la productividad económica y de la capacidad tecnológica. Desarrollos posteriores, relacionados con

la Revolución Industrial, provocaron un segundo cambio comparable en ritmo de crecimiento.

Estos cambios en el ritmo de crecimiento tienen importantes consecuencias. Hace unos cien mil años, a principios de la prehistoria humana (u homínida), el crecimiento era tan lento que se necesitaron casi un millón de años para aumentar la capacidad humana productiva lo suficiente como para sostener a un millón de personas adicional a nivel de subsistencia. Alrededor del 5000 a.C., tras la Revolución Agrícola, el ritmo de crecimiento había aumentado hasta el punto de que la misma cantidad de crecimiento sólo requirió dos siglos. Hoy en día, después de la Revolución Industrial, la economía mundial crece de media esa cantidad cada noventa minutos.[1]

Incluso el actual ritmo de crecimiento producirá resultados impresionantes si se mantiene durante un período de tiempo moderadamente prolongado. Si la economía mundial sigue creciendo al mismo ritmo que viene haciéndolo durante los últimos cincuenta años, el mundo será 4,8 veces más rico en el 2050, y unas 34 veces más rico en el 2100 de lo que es actualmente.[2]

Sin embargo, la perspectiva de continuar por el camino del crecimiento exponencial constante palidece en comparación con lo que pasaría si el mundo experimentase otro cambio en su *ritmo de crecimiento,* de magnitud comparable a los relacionados con la Revolución Agrícola y la Revolución Industrial. El economista Robin Hanson estima, basándose en datos económicos y demográficos históricos, que una economía mundial característica que se multiplica por dos en la sociedad de cazadores-recolectores del Pleistoceno cada 224.000 años; en la sociedad agrícola, cada 909 años; y en la sociedad industrial, cada 6,3 años.[3] (En el modelo de Hanson, la época actual es una combinación de los modos de crecimiento agrícola e industrial; la economía mundial, en su conjunto, aún no está creciendo al ritmo de duplicación de 6,3 años). Si tuviera lugar otra transición similar hacia un patrón de crecimiento distinto, y si fuese de una magnitud similar a los dos anteriores, daría como resultado un nuevo régimen de crecimiento en el que la economía mundial se duplicaría de tamaño aproximadamente cada dos semanas.

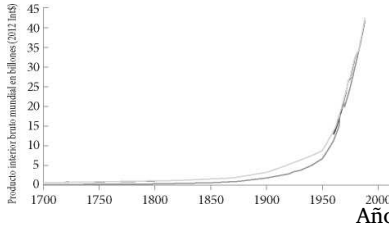
Dicho ritmo de crecimiento parece fantasioso por lo que actualmente sabemos. Es posible que los observadores de épocas anteriores hayan considerado igualmente absurdo suponer que la economía mundial llegaría a duplicarse varias veces a lo largo de una sola vida. No obstante, ésa es la extraordinaria condición que ahora consideramos ordinaria.

La idea de una singularidad tecnológica inminente ya se ha popularizado bastante, comenzando con el ensayo fundacional de Vernor Vinge y continuando con los escritos de Ray Kurzweil y otros.[4]

Sin embargo, el término “singularidad” se ha utilizado de forma confusa en muchos y diversos sentidos y ha adquirido un aura impura (casi mi- lenarista) de connotaciones técnico-utópicas.[5] Dado que la mayoría de estos significados y connotaciones son irrelevantes para nuestras explicaciones, podemos ganar en claridad obviando la palabra “singularidad” en favor de una terminología más precisa.

La idea relacionada con la singularidad que aquí nos interesa es la posibilidad de una *explosión de inteligencia,* especialmente la perspectiva de una superinteligencia artificial. Tal vez haya personas convencidas, debido a diagramas de crecimiento como los de la Figura 1, de que otro cambio drástico en el modo de crecimiento, comparable al de la Revolución Agrícola o la Revolución Industrial, está a la vuelta de la esquina. Estas personas puede entonces darse cuenta de que es difícil concebir un escenario en

(a) 45
40
35
30
I 25
%
20
15
| 10
5
I 0



-8000 -6000 -4000 -2000
Año

el que el tiempo de duplicación de la economía mundial se reduzca a meras semanas que no conlleve la creación de mentes que sean mucho más rápidas y eficientes que las de tipo biológico familiar. Sin embargo, el hecho de tomarse en serio la perspectiva de una revolución en inteligencia artificial no tiene por qué depender de ejercicios de ajustes de curvas, ni de extrapolaciones a partir de antiguos crecimientos económicos. Como veremos, hay razones más convincentes para tomársela en serio.

Grandes expectativas

Que las máquinas se equiparen a los seres humanos en inteligencia general —es decir, en tener sentido común y una capacidad eficiente para aprender, razonar y planificar para afrontar complejos retos de procesamiento de información a lo largo de un amplio rango de ámbitos naturales y abstractos— ha sido algo esperado desde la invención de los ordenadores, en la década de los 40 del siglo XX. En aquella época, la llegada de este tipo de máquinas a menudo solía preverse para dentro de veinte años.[7]

Desde entonces, la fecha prevista de llegada ha ido retrocediendo a un ritmo de un año por cada año; por lo que hoy, los futuristas que se preocupan por la posibilidad de la inteligencia artificial general aún suelen creer que las máquinas inteligentes están a un par de décadas de distancia.[8]

Dos décadas es un punto óptimo para los pronosticadores de cambios radicales: lo suficientemente cerca como para que llame la atención y sea relevante, pero lo bastante lejos como para suponer que una serie de innovaciones, actualmente sólo vagamente imaginables, pudieran tener lugar. Contrastemos esto con plazos temporales más breves: la mayor parte de las tecnologías que tendrán un gran impacto en el mundo en cinco o diez años a partir de ahora ya están disponibles para su uso limitado, mientras que las tecnologías que darán forma al mundo en menos de quince años probablemente existen como prototipos de laboratorio. Veinte años puede también ser un período de tiempo cercano a la duración de la carrera de adivino, limitando el riesgo que su reputación podría sufrir por una predicción arriesgada.

Sin embargo, del hecho de que algunos individuos hayan predicho erróneamente la inteligencia artificial en el pasado, no se sigue que la IA sea imposible o que nunca vaya a desarrollarse.[9] La principal razón por la que el progreso ha sido más lento de lo esperado es que las dificultades técnicas en la construcción de máquinas inteligentes han sido superiores de lo que los pioneros previeron. Pero esto tan sólo pone de manifiesto la magnitud de las dificultades y lo lejos que estamos de superarlas. En ocasiones, un problema que en un principio parece muy complicado resulta tener una solución sorprendentemente simple (aunque seguramente sea más común lo contrario).

En el próximo capítulo, examinaremos diferentes caminos que pueden conducir a la inteligencia artificial de nivel humano. Pero permítasenos señalar desde el principio que, a pesar de las numerosas paradas que pueda haber entre el momento actual y la inteligencia artificial de nivel humano, ésta no es el destino final. La siguiente parada, a escasa distancia a lo largo de la misma senda, es la inteligencia artificial de nivel sobrehumano. El tren podría no detenerse ni decelerar en la estación *Villahumana*. Es probable que la atraviese como una flecha.

El matemático I. J. Good, que trabajó como jefe de estadística en el equipo de decodificadores de Alan Turing durante la Segunda Guerra Mundial, ha sido tal vez el primero en enunciar los aspectos esenciales de este escenario. En un pasaje muy citado de 1965, escribió:

Definamos una máquina ultrainteligente como aquella que puede superar con creces

todas las actividades intelectuales de cualquier hombre por muy listo que sea. Puesto que el diseño de máquinas es una de esas actividades intelectuales, una máquina ultrainteligente podría diseñar máquinas incluso mejores; entonces habría, sin duda, una "explosión de inteligencia", y la inteligencia humana quedaría muy atrás. Por ello, la primera máquina ultrainteligente es el último invento que el hombre necesita crear, contando con que la máquina sea lo suficientemente dócil como para decirnos cómo mantenerla bajo control.[10]

Ahora puede ya parecer obvio que hay grandes riesgos existenciales relacionados con una explosión de inteligencia, y que esta perspectiva, por tanto, debería ser examinada con la mayor seriedad, aunque se supiera (lo cual no es así) que sólo hay una probabilidad moderadamente pequeña de que ocurra. Sin embargo, los pioneros de la inteligencia artificial, a pesar de su creencia en la inminente llegada de la IA de nivel humano, en su mayoría no contemplaron la posibilidad de una IA superior a la humana. Es como si su capacidad especulativa hubiera quedado tan agotada concibiendo la posibilidad radical de máquinas que pudieran alcanzar la inteligencia humana, que no llegaron a su corolario —que las máquinas subsecuentemente llegarían a ser su- perinteligentes.

Los pioneros de la IA, en su mayoría, no contemplaron la posibilidad de que su empresa pudiera implicar ningún riesgo.[11] No hablaron —ni pensaron seriamente— sobre ningún problema de seguridad o cuestión ética relacionada con la creación de mentes artificiales y posibles soberanos informáticos: una laguna que asombra incluso en el contexto del poco impresionante estándar de la época para asuntos de valoración crítica de la tecnología.[12] Debemos tener la esperanza de que, en el momento en que esa empresa sea realizable, tendremos no sólo la competencia tecnológica para desencadenar una explosión de inteligencia sino también el nivel superior de maestría necesario para hacer que la detonación no sea mortal.

Pero, antes de afrontar lo que se nos presenta delante, será útil echar un rápido vistazo a la historia de la inteligencia artificial hasta la fecha.

Epocas de esperanza y desesperación

En el verano de 1956, en el Dartmouth College, diez científicos que compartían su interés por las redes neuronales, la teoría de autómatas y el estudio de la inteligencia, se reunieron para celebrar un taller de seis semanas de duración. Este Dartmouth Summer Project suele considerarse el punto de partida de la inteligencia artificial como campo de investigación. Muchos de los participantes serían reconocidos posteriormente como pioneros. La visión optimista entre los delegados se refleja en la propuesta presentada a la Fundación

Rockefeller, que aportó los fondos para el evento:

> Proponemos que diez hombres, durante dos meses, lleven a cabo un estudio sobre la inteligencia artificial... El estudio partirá de la conjetura de que cualquier aspecto del aprendizaje, o cualquier otra característica de la inteligencia puede, en principio, ser descrita con tanta precisión como para que pueda fabricarse una máquina que la simule. Se intentará descubrir cómo fabricar máquinas que usen lenguaje, que formen abstracciones y conceptos, que resuelvan problemas hasta ahora reservados a los seres humanos y que se mejoren a sí mismas. Creemos que puede conseguirse un avance significativo en uno o más de estos problemas si un grupo de científicos cuidadosamente seleccionados trabajan en ello juntos durante un verano.

En las seis décadas transcurridas desde este audaz comienzo, el campo de la inteligencia artificial ha pasado por períodos de entusiasmo generalizado y altas expectativas alternando con períodos de retroceso y decepción.

El primer período de entusiasmo, que comenzó con la reunión de Dartmouth, fue descrito más tarde por John McCarthy (el principal organizador del evento) como la era del "¡mira mamá, sin manos!" Durante esta primera época, los investigadores construyeron sistemas diseñados para refutar afirmaciones del tipo "¡ninguna máquina podría jamás hacer *X!*". Tales afirmaciones escépticas eran comunes por aquella época. Para contrarrestarlas, los investigadores de la IA crearon pequeños sistemas que lograban *X* en un "micromundo" (un ámbito limitado y bien definido que permitía una versión reducida del desempeño que se deseaba demostrar), con lo que aportaron una prueba conceptual y mostraron que *X* podría, en principio, ser realizada por una máquina. Uno de estos primeros sistemas, el Teórico Lógico, fue capaz de demostrar la mayor parte de los teoremas del segundo capítulo de los *Principia Mathematica* de Whitehead y Russell, e incluso proporcionó una prueba que era mucho más elegante que la original, con lo que desacreditó la idea de que las máquinas "sólo podían pensar numéricamente" y demostró que las máquinas también eran capaces de hacer deducciones e idear pruebas lógicas.[13] Un programa posterior, el Solucionador General de Problemas, podía resolver, en principio, una amplia gama de problemas especificados formalmente.[14] También se elaboraron programas que podían resolver problemas de cálculo típicos de los cursos de primer año de universidad, problemas de analogía visual como los que aparecen en algunos tests de CI, y sencillos problemas de álgebra verbal escrita.[15] El Robot Shakey (llamado así debido a su tendencia a temblar durante su funcionamiento) demostró cómo el razonamiento lógico podía integrarse con la percepción y utilizarse para planificar y controlar la actividad física.[16] El programa ELIZA demostró que un ordenador podía hacerse pasar por un psicoterapeuta Roge- riano.[17] A mediados de los años setenta, el programa SHRDLU mostró cómo la simulación de un brazo robótico, en un mundo simulado de bloques geométricos,

podía seguir instrucciones y responder a las preguntas tecleadas en inglés por un usuario.[18] En las décadas posteriores, se crearon sistemas que demostraron que las máquinas podían componer música al estilo de diversos compositores clásicos, superar a los médicos novatos en determinadas tareas de diagnóstico clínico, conducir coches de forma autónoma y hacer inventos patentables.[19] Ha habido incluso una IA que hizo chistes originales.[20] (Su nivel de humor no era muy alto —"¿Qué se obtiene cuando se cruza un óptico con un *objeto mental?* Una /-dea"— pero los niños expresaron repetidamente que sus bromas les resultaban consistentemente entretenidas).

Los métodos que produjeron éxitos en los primeros sistemas de demostración solían también ser difíciles de extender a una variedad más amplia de problemas o a ejemplos de problemas más complicados. Una razón para ello es la "explosión combinatoria" de posibilidades que deben ser exploradas para los métodos basados en algo parecido a la búsqueda exhaustiva. Estos métodos funcionan bien para los casos sencillos de un problema, pero fracasan cuando las cosas se ponen un poco más complicadas. Por ejemplo, para demostrar un teorema de unas cinco líneas de extensión en un sistema deductivo con una regla de inferencia y 5 axiomas, podríamos limitarnos a enumerar las 3.125 posibles combinaciones y comprobar cada una de ellas para ver si alguna proporciona la conclusión deseada. La búsqueda exhaustiva también funcionaría para demostraciones de seis o siete líneas. Pero a medida que la tarea se vuelve más difícil, el método de búsqueda exhaustiva pronto genera problemas. Demostrar un teorema de 50 líneas no requiere un tiempo diez veces mayor que otro de cinco líneas: en lugar de eso, si utilizamos la búsqueda exhaustiva, requiere rastrear 5^{50} « 8,9 x 10^{34} posibles secuencias, lo cual es computacionalmente inviable, incluso para los superordenadores más rápidos.

Para superar la explosión combinatoria, se necesita algoritmos que aprovechen la estructura del ámbito objetivo y que se aprovechen de los conocimientos previos mediante el uso de búsquedas heurísticas, planificación y representaciones abstractas flexibles; unas capacidades que estaban pobremente desarrolladas en los primeros sistemas de IA. El rendimiento de estos primeros sistemas también sufrió por los deficientes métodos para manejar incertidumbre, por la dependencia de representaciones frágiles y de poca base, por la escasez de datos, y por las graves limitaciones de hardware en cuanto a capacidad de memoria y velocidad de procesamiento. A mediados de la década de los 70 del siglo XX, hubo una creciente toma de conciencia de estos problemas. La comprensión de que muchos proyectos de IA nunca podrían cumplir sus promesas iniciales condujo a la aparición del primer "invierno de la IA": un período de retraimiento, durante el cual

disminuyeron los fondos y aumentó el escepticismo, y la IA se pasó de moda.

Una nueva primavera llegó a principios de la década de los 80, cuando Japón lanzó su Proyecto de Sistemas Computacionales de Quinta Generación, una organización público-privada bien financiada que pretendía superar los logros anteriores desarrollando una masiva arquitectura de computación paralelizada que serviría de plataforma para la inteligencia artificial. Esto coincidió con el punto álgido de fascinación por el "milagro económico de posguerra" japonés, un período en el que los líderes gubernamentales y empresariales occidentales buscaban ansiosamente descubrir la fórmula del éxito económico de Japón, con la esperanza de replicar las mismas recetas mágicas en casa. Cuando Japón decidió invertir grandes cantidades en IA, otros países siguieron su ejemplo.

Los años siguientes fueron testigos de una gran proliferación de *sistemas expertos.* Diseñados como herramientas de apoyo a la toma de decisiones, los sistemas expertos son programas basados en reglas que efectuaban sencillas inferencias a partir del conocimiento de hechos básicos, que habían sido obtenidos de expertos humanos en ese ámbito y cuidadosamente codificados a mano en lenguaje formal. Se construyeron cientos de estos sistemas expertos. Sin embargo, los sistemas más pequeños conllevaban escasos beneficios, y los más grandes resultaron caros de desarrollar, validar y actualizar, y generalmente eran difíciles de manejar. Era poco práctico adquirir un ordenador independiente para la tarea de ejecutar un solo programa. A finales de la década de los 80 del siglo XX, esta temporada de crecimiento también se había extinguido.

El Proyecto de Quinta Generación no cumplió con sus objetivos, lo mismo que le sucedió a sus imitadores de Estados Unidos y Europa. Un segundo invierno de la IA sobrevino. En este punto, un crítico podría haber lamentado, con toda justicia, que "la historia de la investigación en inteligencia artificial, hasta la fecha, había consistido en éxitos muy limitados en áreas concretas, seguidos inmediatamente de fracasos al intentar alcanzar los objetivos más amplios a los que esos éxitos iniciales parecían apuntar en un primer momento".[21] Los inversores privados comenzaron a huir de cualquier empresa que llevara la marca de "inteligencia artificial". Incluso entre los académicos y sus financiadores, la "IA" se convirtió en un epíteto poco deseado.[22]

No obstante, el trabajo técnico continuó a buen ritmo, y en la década de los 90 del siglo XX, el segundo invierno de la IA fue gradualmente llegando a su deshielo. El optimismo se reavivó gracias a la introducción de nuevas técnicas, que parecían ofrecer alternativas al paradigma logicista tradicional (a menudo denominado

"inteligencia artificial a la antigua usanza", GOFAI —Good Old-Fashioned Artificial Inte- lligence— como abreviatura), que se había centrado en la manipulación de símbolos de alto nivel y que había llegado a su apogeo en los sistemas expertos de la década de los 80. Las nuevas y populares técnicas, que incluían redes neuronales y algoritmos genéticos, prometían superar algunas de las carencias del paradigma GOFAI, en particular la "fragilidad" que caracterizaba a los programas clásicos de IA (que solían producir completos sinsentidos si los programadores hacían aunque sólo fuera una suposición errónea). Las nuevas técnicas presentaban un rendimiento más orgánico. Por ejemplo, las redes neuronales exhibían la propiedad de "degradación elegante": una pequeña cantidad de daño a una red neuronal solía resultar en una pequeña degradación de su funcionamiento, y no en su derrumbe completo. Y lo que era más importante, las redes neuronales podían aprender de la experiencia, encontrar formas naturales de generalizar a partir de ejemplos y buscar patrones estadísticos ocultos en la información recibida.[23] Esto hizo que las redes fueran buenas en problemas de reconocimiento y clasificación de patrones. Por ejemplo, al entrenar una red neuronal en un conjunto de datos de señales de sónar, se le podía enseñar a distinguir los perfiles acústicos de submarinos, minas y vida marina con mayor precisión que a los expertos humanos —y esto se podía hacer sin que nadie tuviera que averiguar exactamente de antemano cómo debían definirse las categorías o cómo tenían que ponderarse las diferentes características.

Aunque se conocían modelos sencillos de redes neuronales desde finales de la década de los 50 del siglo XX, este campo tuvo un renacimiento después de la introducción del algoritmo de retropropagación, que posibilitó entrenar redes neuronales de capas múltiples.[24] Este tipo de redes de capas múltiples, que tienen una o más capas de neuronas ("ocultas") intermedias entre las capas de entrada y de salida, pueden aprender una gama mucho más amplia de funciones que sus predecesoras más simples.[25] En combinación con los ordenadores cada vez más potentes que había disponibles, estas mejoras algorítmicas permitieron a los ingenieros construir redes neuronales lo suficientemente buenas como para ser útiles en sentido práctico en muchas aplicaciones.

Las cualidades de las redes neuronales similares a las del cerebro contrastaban favorablemente con el rendimiento rigurosamente lógico pero frágil de los tradicionales sistemas GOFAI basados en reglas —lo suficiente como para inspirar la creación de un nuevo "ismo", el *conexionismo,* que hacía hincapié en la importancia del procesamiento sub-simbólico masivo en paralelo. Desde entonces se han publicado más de 150.000 artículos académicos sobre las redes neuronales artificiales, y continúan siendo una perspectiva importante en

aprendizaje artificial.

Los métodos basados en la evolución, como por ejemplo los algoritmos genéticos y la programación genética, constituyen otro enfoque cuya emergencia ha ayudado a poner fin al segundo invierno de la IA. Tal vez tuvo un impacto académico menor que las redes neuronales, pero fue muy popular. En los modelos evolutivos, se mantiene una población de soluciones candidatas (que pueden ser estructuras de datos o programas), y las nuevas soluciones candidatas se generan aleatoriamente mediante mutaciones o variantes que recombinan la población existente. Periódicamente, se reduce la población mediante la aplicación de un criterio de selección (una función de aptitud), que propicia que en la próxima generación sobrevivan sólo los mejores candidatos. Repetida durante miles de generaciones, la calidad media de las soluciones del grupo de candidatos aumenta gradualmente. Cuando funciona, este tipo de algoritmo puede producir soluciones eficientes para una amplia gama de problemas —soluciones que pueden ser sorprendentemente novedosas y poco intuitivas, normalmente más parecidas a las estructuras naturales que cualquier cosa diseñada por un ingeniero humano. Y, en principio, esto puede ocurrir sin más necesidad de intervención humana que la especificación inicial de la función de aptitud, que habitualmente es muy simple. En la práctica, sin embargo, conseguir métodos evolutivos que funcionen bien requiere habilidad e ingenio, sobre todo a la hora de elaborar un buen formato de representación. Sin un procedimiento eficiente para codificar soluciones candidatas (un lenguaje genético que coincida con la estructura latente del ámbito objetivo), la propia búsqueda de la evolución tiende a deambular eternamente en un vasto espacio de búsqueda o queda atrapada en un punto local óptimo. Aunque se encuentre un buen formato de representación, la evolución es computacionalmente exigente, y suele ser derrotada por la explosión combinatoria.

Las redes neuronales y los algoritmos genéticos son ejemplos de métodos que despertaron el entusiasmo en la década de los 90 del siglo XX, al surgir para ofrecer alternativas al paradigma GOFAI, que se había estancado. Pero no es nuestra intención cantar las alabanzas de estos dos métodos, ni elevarlos por encima de las muchas otras técnicas de aprendizaje artificial. De hecho, uno de los principales desarrollos teóricos de los últimos veinte años ha sido una comprensión más clara de cómo técnicas superficialmente diferentes pueden entenderse como casos especiales dentro de un marco matemático común. Por ejemplo, muchos tipos de redes neu- ronales artificiales pueden considerarse clasificadores que realizan un tipo particular de cálculo estadístico (estimación de máxima

probabilidad).[26] Esta perspectiva permite a las redes neuronales compararse con una clase más amplia de algoritmos para aprender a clasificar a partir de ejemplos: “árboles de decisiones”, “modelos de regresión logística”, “vectores de máquinas de apoyo”, “clasificadores bayesianos ingenuos”, “regresión a los ˆ-vecinos más cercanos”, entre otros.[27] De manera similar, los algoritmos genéticos se pueden considerar la realización de una escalada esto- cástica, que es de nuevo un subconjunto de una clase más amplia de algoritmos de optimización. Cada uno de estos algoritmos para la construcción de clasificadores, o para la búsqueda de un espacio de soluciones, tiene su propio perfil de puntos fuertes y débiles que se pueden estudiar matemáticamente. Los algoritmos difieren en los requerimientos relativos a su tiempo de proceso y espacio de memoria, en los sesgos inductivos que presuponen, en la facilidad con que se puede incorporar contenido producido externamente, y en cómo de transparentes son sus procesos internos para un analista humano.

Por tanto, detrás del alboroto alrededor del aprendizaje artificial y de la resolución creativa de problemas, encontramos un conjunto de compensaciones matemáticamente bien especificadas. El ideal es el de un agente bayesiano perfecto, uno que haga un uso probabilísticamente óptimo de la información disponible. Este ideal es inalcanzable, porque, computacionalmente, es demasiado exigente para que sea im- plementado en cualquier ordenador físico (véase el cuadro 1). En consecuencia, se puede considerar la inteligencia artificial como una búsqueda por encontrar atajos: procedimientos para aproximarse flexiblemente al ideal bayesiano sacrificando algunas optimizaciones o generalidades, a la vez que se preserva lo suficiente como para conseguir un alto rendimiento en los ámbitos de auténtico interés.

Un reflejo de esta imagen se puede ver en el trabajo realizado durante el último par de décadas en los modelos gráficos probabilísticos, como las redes bayesianas. Las redes bayesianas proporcionan una forma concisa de representar las relaciones de independencia probabilísticas y condicionales que comparten algún ámbito particular. (Explotar esas relaciones de independencia es esencial para superar la explosión combinatoria, que es un gran problema tanto para la inferencia probabilística como para la deducción lógica). También nos ayudan de manera importante a entender el concepto de causalidad.[28]

Cuadro 1. *Continúa*

La regla de aprendizaje y la regla de decisión en conjunto definen una "noción de optimización" para un agente. (Esencialmente, la misma noción de optimización se ha utilizado ampliamente en la inteligencia artificial, la epistemología, la filosofía de la ciencia, la economía y la estadística[34]). En realidad, es imposible construir un agente así, porque es computacionalmente inviable realizar los cálculos necesarios. Cualquier intento de hacerlo sucumbe a una explosión combinatoria como la descrita en nuestra explicación de la GOFAI. Para comprobar por qué esto es así, consideremos un pequeño subconjunto de todos los mundos posibles: aquellos que constan de un único monitor de ordenador flotando en un vacío sin fin. El monitor tiene 1.000 x 1.000 píxeles, cada uno de los cuales está constantemente encendido o apagado. Incluso este subconjunto de mundos posibles es enormemente grande: los $2^{(1000 \times 1000)}$ posibles estados del monitor superan en número a todos los cálculos que se puede esperar que tengan lugar en el universo observable. Por tanto, ni siquiera podríamos enumerar todos los mundos posibles en este diminuto subconjunto de todos los mundos posibles, y mucho menos realizar cálculos individuales más elaborados de cada uno de ellos.

Las nociones de optimización pueden tener interés teórico, aunque sean físicamente irrealizables. Nos ofrecen un estándar desde el que juzgar las aproximaciones heurísticas, y a veces podemos razonar acerca de lo que un agente óptimo haría en algún caso especial. Nos encontraremos con algunas nociones alternativas de optimización para agentes artificiales en el Capítulo 12.

J

Una de las ventajas de relacionar los problemas de aprendizaje de ámbito específico con el problema general de la inferencia bayesiana es que los nuevos algoritmos que hacen más eficiente la inferencia bayesiana producirán mejoras inmediatas en muchas áreas diferentes. Los avances en las técnicas de aproximación Monte Carlo, por ejemplo, se aplican directamente a la visión artificial, la robótica y la genética computacional. Otra ventaja es que permite a investigadores de diferentes disciplinas agrupar más fácilmente sus hallazgos. Los modelos gráficos y las estadísticas baye- sianas se han convertido en un objeto común a la investigación de muchos campos, incluyendo el aprendizaje artificial, la física estadística, la bioinformática, la optimización combinatoria y la teoría de la comunicación.[35] Una buena cantidad de los progresos recientes en aprendizaje artificial procede de la incorporación de resultados formales obtenidos originalmente en otros campos académicos. (Las aplicaciones de aprendizaje artificial también se han beneficiado enormemente de ordenadores más rápidos y de una mayor disponibilidad de grandes cantidades de datos).

Estado de la técnica

La inteligencia artificial ya supera a la inteligencia humana en muchos ámbitos. La tabla 1 refleja el estado de los ordenadores dedicados a juegos, y demuestra que las IAs ya ganan a los campeones humanos en una amplia serie de juegos.[36]

El programa de damas de Arthur Samuel, escrito originalmente en 1952 y posteriormente mejorado (la versión de 1955 incorporó el aprendizaje artificial) se convierte en el primer programa que aprende a jugar a un juego mejor que su creador[37] En 1994, el programa CHINOOK vence al vigente campeón humano, lo cual supone la primera vez que un programa gana un campeonato mundial oficial en un juego de habilidad. En 2002, Jonathan Schaeffer y su equipo "resuelven" el juego de las damas, es decir diseñan un programa que hace siempre el mejor movimiento posible (combinando una búsqueda alfa-beta con una base de datos de 39 billones de finales). El juego perfecto por parte de ambos bandos conduce a un empate.[38]

1979: el programa de backgammon BKG, de Hans Berliner derrota al campeón del mundo — el primer programa de ordenador que venció (en un encuentro de exhibición) a un campeón mundial en algún tipo de juego—, aunque Berliner posteriormente atribuye la victoria a la suerte en las tiradas de dados.[39] 1992: el programa de backgammon TD-Gammon de Gerry Tesauro alcanza un nivel de campeonato, utilizando un aprendizaje diferencial temporal (una forma de aprendizaje por refuerzo), y juega contra sí mismo repetidamente, para mejorar[40] En los años posteriores, los programas de backgammon han superado con creces a los mejores jugadores humanos.[41]

En 1981 y 1982, el programa Eurisko de Douglas Lenat gana el campeonato estadounidense de Traveller (un futurista juego de guerra naval), lo cual lleva a cambiar las reglas para evitar sus estrategias poco ortodoxas.[43] Eurisko utilizaba la heurística para diseñar su flota, y también utilizaba la heurística para modificar su heurística.

Nivel sobrehumano: 1977: el programa Logistello gana todas las partidas de un encuentro a seis partidas contra el campeón del mundo, Takeshi Murakami.[44]

Nivel sobrehumano: 1997: Deep Blue vence al campeón del mundo de ajedrez, Gary Kasparov. Kasparov afirma haber detectado destellos de verdadera inteligencia y creatividad en algunos de los movimientos del ordenador[45] Desde entonces, los motores de ajedrez han seguido mejorando.[46]

Nivel experto 1999: el programa de resolución de crucigramas Proverb, supera al solucionador de crucigramas medio.[47]

Nivel sobrehumano	
Scrabble	2002: los programas dedicados a jugar al Scrabble superan a los mejores jugadores humanos.[49]
Bridge	2009: los programas dedicados a jugar al bridge alcanzan el mismo nivel que los mejores jugadores humanos.[50]
Jeopardy!	2011: el programa Watson, de IBM, derrota a los dos mejores campeones humanos de *Jeopardy!* de todos los tiempos, Ken Jennings y Brad Rutter[51] *Jeopardy!* es un juego televisivo, con preguntas misceláneas sobre historia, literatura, deportes, geografía, cultura pop, ciencia y otros temas. Las preguntas se presentan con pistas y suelen incluir juegos de palabras.
Póker	Los programas de póker informáticos están un poco por debajo de los mejores humanos de la modalidad Texas hold'em, pero juegan a nivel sobrehumano a algunas variantes de póker[52]
Carta blanca	Carta blanca evolucionadas usando algoritmos genéticos producen un solucionador para el juego de solitario de Carta Blanca (que en su forma generalizada es NP- completo) que es capaz de vencer a jugadores humanos de alto nivel.[53]
Go	En 2012, la serie Zen de programas dedicados a jugar al go ha alcanzado un rango de sexto dan en partidas rápidas (el nivel de un jugador aficionado muy fuerte) utilizando la búsqueda de árbol de variantes de Monte Carlo y técnicas de aprendizaje artificial.[54] Los programas que juegan al go han mejorado a un ritmo de más o menos 1 dan/año en los últimos años. Si este ritmo de mejora continua, podrían vencer al campeón del mundo humano en aproximadamente una década.

Estos logros tal vez no parezcan impresionantes hoy en día. Pero esto se debe a que nuestros estándares de lo que es impresionante se van adaptando a los avances que se van realizando. Jugar de manera experta al ajedrez, por ejemplo, era considerado el culmen del intelecto humano. En opinión de varios expertos de finales de los cincuenta: “Si se pudiera diseñar de manera exitosa una máquina de ajedrez, parecería que habríamos penetrado en el núcleo central de los esfuerzos intelectuales humanos”.[55] Esto ya no parece ser así. Uno simpatiza con John McCarthy, quien se lamentaba: “En cuanto funciona, ya nadie lo llama LA”.[56]

Sin embargo, hay un sentido importante en el que la IA de ajedrez fue un triunfo menor del que muchos esperaban. Antes se suponía, tal vez no sin razón, que para que un ordenador jugase al ajedrez a nivel de gran maestro, tendría que estar dotado de un alto grado de inteligencia *general.* [57] Se podía pensar, por ejemplo, que el gran juego del ajedrez requeriría ser capaz de aprender conceptos abstractos, pensar inteligentemente sobre estrategia, elaborar planes flexibles, realizar una amplia gama de ingeniosas deducciones lógicas y tal vez incluso de representarse el pensamiento del rival. No fue así. Resultó posible construir un programa de ajedrez perfectamente adecuado en torno a un algoritmo de objetivo específico.[58] Cuando se implementa en los rápidos procesadores disponibles a finales del siglo XX, da lugar a un nivel de juego muy fuerte. Pero una IA construida de esta manera es una cosa limitada. Juega al ajedrez; pero no puede hacer otra cosa.[59]

En otros ámbitos, las soluciones han resultado ser *más* complicadas de lo previsto inicialmente, y el progreso más lento. El científico computacional Donald Knuth se sorprendió de que "la IA haya tenido ya éxito haciendo básicamente todo lo que requiere «pensar», pero haya fracasado en lo que la mayoría de la gente y los animales hacen «sin pensar» —eso, de algún modo, ¡es mucho más difícil!".[60] Analizar escenas visuales, reconocer objetos o controlar el comportamiento de un robot que interactúa con un entorno natural ha demostrado ser todo un reto. Sin embargo, se ha conseguido una buena cantidad de progreso y se sigue progresando, ayudado por constantes progresos en el hardware.

El sentido común y la comprensión del lenguaje natural también han resultado ser difíciles. En la actualidad, se suele pensar que alcanzar un rendimiento completamente humano en estas tareas es un problema de "IA completo", queriendo con esto decir que solucionar estos problemas es esencialmente equivalente a la dificultad de crear máquinas de inteligencia general de nivel humano.[61] En otras palabras, si alguien *fuera* a tener éxito creando una IA que pudiera entender el lenguaje natural tal como lo hace un adulto humano, también habría ya tenido éxito, con toda probabilidad, en la tarea de crear una IA que pudiese hacer todo lo que una inteligencia humana puede hacer, o estaría muy cerca de una capacidad general de esa clase.[62]

La maestría en ajedrez resultó ser alcanzable por medio de un algoritmo sorprendentemente simple. Resulta tentador especular que otras capacidades —tales como la capacidad de razonamiento general o alguna habilidad clave implicada en programación— podrían ser igualmente alcanzables a través de algún algoritmo sorprendentemente simple. El hecho de que se alcance un mejor

rendimiento mediante un mecanismo complicado no significa que no exista un mecanismo simple que pueda hacer el trabajo igual o mejor. Podría suceder tan sólo que nadie hubiese encontrado esa alternativa más simple. El sistema de Ptolomeo (con la Tierra en el centro y el Sol, la Luna, los planetas y las estrellas orbitando en torno a ella) representó el estado de la técnica en astronomía durante más de mil años, y su exactitud predictiva fue mejorando con el paso de los siglos mediante una gradual complicación del modelo: añadiendo epiciclos sobre epiciclos a los movimientos celestes postulados. Después, todo el sistema fue derrocado por la teoría heliocéntrica de Copérnico, que era más simple, y —aunque sólo después de una mayor elaboración por parte de Kepler— más precisa en lo referente a las predicciones.[63]

Los métodos de inteligencia artificial se utilizan ahora en más áreas de las que tendría sentido revisar aquí, pero mencionar una muestra dará una idea de la amplitud de las aplicaciones. Aparte de las IA dedicadas a juegos reflejadas en la tabla 1, hay audífonos con algoritmos que filtran el ruido ambiental; buscadores de rutas que muestran mapas y ofrecen consejos de navegación a los conductores; sistemas de recomendación que sugieren libros y álbumes de música basándose en las compras y clasificaciones anteriores de un usuario; y sistemas de asistencia para decisiones médicas que ayudan a los médicos a diagnosticar el cáncer de mama, que recomiendan planes de tratamiento, y que ayudan en la interpretación de electrocardiogramas. Hay mascotas robóticas y robots de limpieza, robots que cortan el césped, robots de rescate, robots quirúrgicos y más de un millón de robots industriales.[64] La población mundial de robots supera los 10 millones.[65]

El moderno reconocimiento de voz, basado en técnicas estadísticas como los modelos ocultos de Markov, se ha vuelto lo suficientemente preciso para su uso práctico (algunos fragmentos de este libro se elaboraron con la ayuda de un programa de reconocimiento de voz). Asistentes digitales personales, tales como Siri de Apple, responden a órdenes verbales y pueden responder preguntas sencillas y ejecutar órdenes. El reconocimiento óptico de caracteres de texto manuscrito y mecanografiado se utiliza rutinariamente en aplicaciones como la clasificación del correo y la digitalización de documentos antiguos.[66]

La traducción automática sigue siendo imperfecta pero es suficientemente buena para muchas aplicaciones. Los primeros sistemas utilizaban el enfoque GOFAI, con gramáticas de codificación manual que tuvieron que ser desarrolladas desde cero por lingüistas cualificados en cada idioma. Los nuevos sistemas utilizan técnicas estadísticas de aprendizaje artificial que construyen automáticamente modelos estadísticos a partir de los patrones de uso observados. La máquina deduce los parámetros de estos modelos mediante un análisis

bilingüe del texto. Este enfoque prescinde de los lingüistas: los programadores que construyen estos sistemas ni siquiera tienen por qué hablar las lenguas con las que trabajan.[67]

El reconocimiento facial ha mejorado tanto en los últimos años que ahora se utiliza en los pasos fronterizos automatizados de Europa y Australia. El Departamento de Estado de los Estados Unidos cuenta con un sistema de reconocimiento facial de más de 75 millones de fotografías para la tramitación de visados. Los sistemas de vigilancia utilizan una IA cada vez más sofisticada, además de tecnología de análisis de datos para analizar la voz, vídeos o textos, un buen número de los cuales se obtienen de los medios de comunicación electrónicos de todo el mundo y se almacenan en enormes bases de datos.

La demostración de teoremas y la resolución de ecuaciones están actualmente tan asentadas que apenas se siguen considerando como IA. En los programas informáticos científicos se incluyen solucionadores de ecuaciones como Mathematica. Los métodos de verificación formal, incluyendo demostradores de teoremas automatizados, suelen ser utilizados por fabricantes de chips de forma rutinaria para verificar el comportamiento de los diseños de circuitos antes de su producción.

Los organismos militares y de inteligencia de los Estados Unidos han liderado la senda que ha llevado al despliegue a gran escala de robots desactivadores de bombas, drones de vigilancia y ataque, y de otros vehículos no tripulados. Éstos todavía dependen principalmente del control remoto por parte de operadores humanos, pero se está trabajando para ampliar sus capacidades autónomas.

La programación inteligente es un ámbito muy importante en el que se ha tenido bastante éxito. La herramienta DART para la planificación y programación de logística automatizada se utilizó en la Operación Tormenta del Desierto, en 1991, hasta el punto de que la DARPA (la Agencia de Proyectos de Investigación Avanzada en Defensa de los Estados Unidos) afirma que esta aplicación, por sí sola, ha compensado sobradamente su inversión por treinta años en IA.[68] Los sistemas de reserva de las aerolíneas utilizan sofisticados sistemas de programación y de establecimiento de precios. Las empresas hacen un amplio uso de técnicas de IA en sus sistemas de control de inventario. También utilizan sistemas automáticos de reserva telefónica y líneas de ayuda asociadas a programas de reconocimiento de voz para guiar a sus desesperados clientes por laberintos de opciones de menú interconectadas.

Numerosos servicios de internet se basan en tecnologías de IA. Hay software encargado de vigilar el tráfico de correo electrónico por todo el mundo, y a pesar de la continua adaptación por parte de quienes envían spam para burlar las contramedidas tomadas contra

ellos, los filtros de spam bayesianos han logrado mantener a raya, hasta cierto punto, la marea de correo basura. Softwares que utilizan componentes de IA son los responsables de aprobar o rechazar automáticamente las transacciones con tarjetas de crédito y vigilar continuamente la actividad de las cuentas para detectar los indicios de uso fraudulento. Los sistemas de recuperación de información también hacen un amplio uso del aprendizaje artificial. El motor de búsqueda de Google es, probablemente, el mayor sistema de IA que se ha construido hasta la fecha.

Ahora bien, hay que insistir en que la diferencia entre la inteligencia artificial y los programas de ordenador en general no está definida. Algunas de las aplicaciones mencionadas anteriormente podrían más bien considerarse aplicaciones genéricas que IA específicamente —aunque esto nos lleve de nuevo a la máxima de McCarthy de que, cuando algo funciona, ya no se le llama IA. Una distinción más relevante para nuestros propósitos es la que existe entre los sistemas que tienen un nivel reducido de capacidad cognitiva (se autodenominen "IA" o no) y los sistemas que tienen capacidad de resolución de problemas de aplicación más generalizada. Prácticamente, todos los sistemas actuales en uso son del primer tipo: de nivel reducido. No obstante, muchos de ellos contienen componentes que también podrían desempeñar un papel en la inteligencia artificial general del futuro, o estar al servicio de su desarrollo —componentes como los clasificadores, los algoritmos de búsqueda, los planificadores, los solucionadores de problemas y las infraestructuras de representación.

Un entorno de alto riesgo y muy competitivo en el que operan actualmente los sistemas de IA es el del mercado financiero global. Las principales compañías inversoras hacen un uso generalizado de sistemas automatizados de intercambio de acciones. Aunque algunos de éstos son sólo formas de automatizar la ejecución de determinadas órdenes de compra o venta emitidas por un gestor de fondos humano, otros conllevan complejas estrategias comerciales que se adaptan a las condiciones cambiantes del mercado. Los sistemas analíticos utilizan diversas técnicas de análisis de datos y series temporales, con el objetivo de buscar patrones y tendencias en los mercados de valores, o para correlacionar los movimientos de precios históricos con variables externas, como palabras clave en los teletipos de los noticieros. Algunos proveedores de noticias financieras venden sistemas de transmisión de noticias que están especialmente formateados para su uso por este tipo de programas de IA. Otros sistemas se especializan en la búsqueda de oportunidades de arbitraje dentro de mercados o entre

Cuadro 2. *Continúa*

previstas en las que sus premisas resultan no ser válidas. El algoritmo sólo hace lo que hace; y, a menos que se trate de un tipo muy especial de algoritmo, no le importa que nos llevemos las manos a la cabeza y nos sorprendamos con horror estupefaciente ante lo inapropiado y absurdo de sus acciones. Éste es un tema que nos volveremos a encontrar Una tercera observación relacionada con el Flash Crash es que, aunque la automatización contribuyó al incidente, también contribuyó a su resolución. La orden de suspensión de la lógica preprogramada, que detuvo el comercio cuando los precios se alejaron de los límites razonables, se estableció para que fuera ejecutada automáticamente porque se había previsto acertadamente que los acontecimientos desencadenantes podrían ocurrir en una escala temporal demasiado rápida como para que los seres humanos respondieran. La necesidad de una función de seguridad preinstalada que se ejecute automáticamente —frente a la confianza en la supervisión humana a tiempo real— otra vez anuncia un tema que será importante en nuestra discusión sobre la inteligencia artificial.[72]

ellos, o en el comercio de alta frecuencia, que busca sacar provecho de los pequeños movimientos de precios que se producen en el transcurso de milisegundos (una escala de tiempo en la que las demoras de comunicación, incluso con señales transmitidas a la velocidad de la luz mediante cables de fibra óptica, se vuelven significativas, haciendo que sea beneficioso colocar los ordenadores cerca de la central). Los comerciantes algorítmicos de alta frecuencia manejan más de la mitad de las transacciones de acciones negociadas en mercados estadounidenses.[69] El comercio algorítmico estuvo implicado en el Flash Crash de 2010 (véase cuadro 2).

Opiniones sobre el futuro de la inteligencia artificial

El progreso en dos grandes frentes —hacia una base estadística y una teoría de la información más sólida para el aprendizaje artificial, por un lado, y hacia el éxito práctico y comercial de varias aplicaciones a problemas concretos o de ámbito específico, por el otro— ha restaurado parte del prestigio perdido a la investigación en IA. No obstante, la historia reciente puede tener un efecto cultural residual en la comunidad de la IA que haga que muchos investigadores punteros no quieran caer en un exceso de ambición. Así Nils Nilsson, uno de los veteranos en el campo, se queja de que sus colegas actuales carezcan de la audacia de espíritu que impulsó a los pioneros de su propia generación:

> La preocupación por la "respetabilidad" ha tenido, creo, un efecto paralizador en algunos investigadores de IA. Los oigo decir cosas como "La IA solía ser criticada por su superficialidad. Ahora que hemos realizado progresos sólidos, no nos arriesguemos a perder nuestra respetabilidad". Un resultado de este conservadurismo es que los esfuerzos se han concentrado en la "IA débil" —la variedad dedicada a proporcionar ayudas al pensamiento humano— alejándose de la "IA fuerte" —la variedad que intenta mecanizar inteligencia humana de nivel humano.[73]

El sentimiento de Nilsson ha sido repetido por varios otros fundadores, entre ellos Marvin Minsky, John McCarthy y Patrick Winston.[74]

Los últimos años han visto un resurgimiento del interés en la IA, que aún podría extenderse a nuevos esfuerzos hacia la inteligencia artificial *general* (lo que Nilsson llama "IA fuerte"). Además de un hardware más rápido, un proyecto actual se beneficiaría de los grandes avances que se han hecho en numerosos sub-campos de la IA, en ingeniería informática de nivel más general, y en ámbitos vecinos, como por ejemplo la neurociencia computacional. Un indicio de la mayor demanda de información y educación de calidad queda demostrado en la respuesta a la oferta por internet de un curso de introducción a la inteligencia artificial en la Universidad de Stanford en el otoño de 2011, organizado por Sebastian Thrun y Peter Norvig. Unos 160.000 estudiantes de todo el mundo se inscribieron para realizarlo (y 23.000 lo finalizaron).[75]

Las opiniones de los expertos sobre el futuro de la IA varían enormemente. No hay acuerdo sobre la sucesión temporal de los acontecimientos ni sobre qué formas podría llegar a adoptar la IA. Las predicciones sobre el futuro desarrollo de la inteligencia artificial, señaló un estudio reciente, "son tan firmes como diversas".[76]

Aunque la distribución actual de estas creencias no se ha medido cuidadosamente, podemos obtener una idea aproximada gracias a varias pequeñas encuestas y observaciones informales. En particular, una serie de encuestas recientes han preguntado a miembros de varias comunidades de expertos sobre cuándo esperan que surja una "inteligencia artificial de nivel humano" (HMLI), definida como "una que pueda desempeñar la mayoría de las profesiones humanas, al menos igual de bien que un ser humano típico".[77] Los resultados se reflejan en la Tabla 2. La muestra combinada dio la siguiente estimación (promedio): 10% de probabilidad de que tenga lugar para el 2022, 50% de probabilidad de que tenga lugar aproximadamente en 2040, y un 90% de probabilidad de que surja para 2075. (A los encuestados se les pidió que basaran sus estimaciones en la premisa de que "la actividad científica humana continúa sin ninguna interrupción negativa importante").

Estas cifras deben tomarse con un grano de sal: el tamaño de las muestras es muy pequeño y no tienen por qué representar

necesariamente a la población general de expertos. No obstante, concuerdan con los resultados de otras encuestas.[78]

Los resultados de la encuesta también concuerdan con unas entrevistas recientemente publicadas donde se preguntaba a dos docenas de investigadores de campos relacionados con la IA. Por ejemplo, Nils Nilsson ha tenido una larga y productiva carrera trabajando en problemas de búsqueda, planificación, representación del conocimiento y robótica; es autor de libros de texto sobre inteligencia artificial; y recientemente finalizó la historia más completa de ese ámbito escrita hasta la fecha.[79] Cuando se le preguntó sobre las fechas de llegada de la inteligencia artificial de nivel humano, ofreció la siguiente opinión:[80]

10% de probabilidad: 2030
50% de probabilidad: 2050
90% de probabilidad: 2100

A juzgar por las transcripciones de las entrevistas publicadas, la distribución de probabilidad del profesor Nilsson parece ser bastante representativa de muchos expertos en esta materia —si bien de nuevo hay que destacar que existe una amplia variedad de opiniones: hay profesionales que son sustancialmente más favorables a la IA y esperan con confianza que haya inteligencia artificial de nivel humano para 2020-40, y otros que creen que nunca va a suceder o que se trata de algo que está indefiniblemente lejos.[82] Además, algunos de los entrevistados consideran que la idea de un "nivel humano" de inteligencia artificial está mal definida o es engañosa, o se muestran reticentes por otros motivos a dejar constancia de una predicción cuantitativa.

Tabla 2. ***¿Cuándo conseguiremos una inteligencia artificial de nivel humano?***[81]

	10%	50%
PT-AI	2023	2080
AGI	2022	2065
EETN	2020	2090
TOPI00	2024	2070
Combinados	2022	2076

Mi propio punto de vista es que las cifras medias de la encuesta de expertos no tienen suficiente capacidad predictiva sobre las fechas de advenimiento tardías. Una probabilidad de un 10% en relación con que no habrá aparecido una inteligencia artificial de nivel humano para el año 2075, y ni siquiera para el 2100 (después de establecer la condición de que "la actividad científica humana proseguiría sin importantes interrupciones negativas") me parece demasiado baja.

Históricamente, los investigadores de la IA no han batido récords prediciendo la velocidad de los avances de su propio campo ni la forma que tales avances tomarían. Por un lado, algunas tareas, como jugar al ajedrez, resultaron alcanzables mediante programas sorprendentemente simples; y los detractores que afirmaban que las máquinas "nunca" serían capaces de hacer esto o aquello se equivocaron en repetidas ocasiones. Por otro lado, los errores más típicos entre los expertos han sido subestimar las dificultades para conseguir que un sistema lleve a cabo tareas del mundo real de manera consistente, y sobreestimar las ventajas de su propio proyecto o técnica favorita.

La encuesta también preguntó otras dos cuestiones de importancia para nuestra investigación. Una inquirió a los encuestados acerca de cuánto tiempo pensaban que tardaría en llegar la superinteligencia asumiendo que primero se alcanzaran máquinas de nivel humano. Los resultados están en la Tabla 3.

Otra pregunta inquirió cuál pensaban que sería el impacto general y a largo plazo para la humanidad, de alcanzar la inteligencia artificial de nivel humano. Las respuestas están resumidas en la figura 2.

De nuevo mis propias opiniones difieren un poco de las opiniones expresadas en la encuesta. Yo asigno una probabilidad más alta a que la superinteligencia surja poco después de la inteligencia artificial de nivel humano. También tengo una perspectiva más polarizada sobre las consecuencias, pues pienso que los resultados muy buenos o muy malos son algo más probables que los resultados más equilibrados. Las razones para esto se aclararán más adelante en el libro.

Table 3. ¿Cuándo se tardará desde nivel humano hasta la superinteligencia?

	30 años después de la inteligencia artificial de nivel humano	2 años después de la inteligencia artificial de nivel humano
TOPI00	50%	50%
Combinados	75%	75%

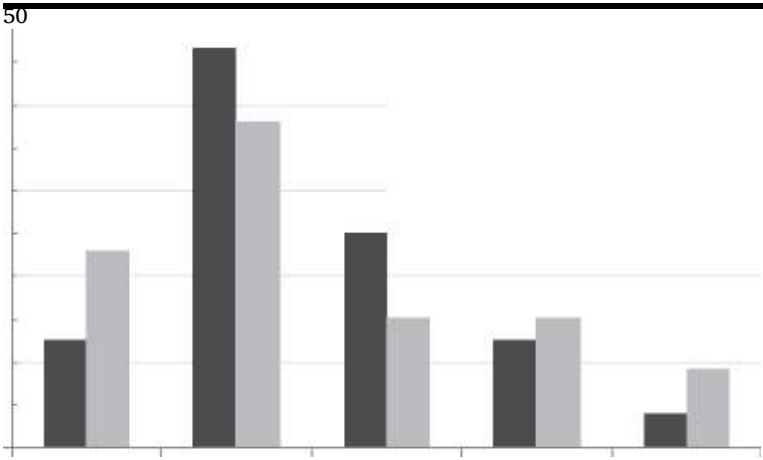


Figura 2. Impacto global a largo plazo de la inteligencia artificial de nivel humano.[8]

Las muestras pequeñas, los sesgos de selección y —sobre todo— la falta de fiabilidad inherente a las opiniones personales implican que no hay que tomarse muy en serio estas encuestas y entrevistas a expertos. No nos permiten sacar ninguna conclusión en firme. Pero sí apuntan a una conclusión débil. Sugieren que (al menos a falta de mejores datos o análisis) puede ser razonable creer que hay probabilidades bastante importantes de que la inteligencia artificial de nivel humano sea desarrollada para mediados de siglo, y que hay una probabilidad nada desdeñable de que sea desarrollada considerablemente antes o mucho más tarde; que tal vez muy poco después de ese momento podría dar como resultado superinteligencia; y que hay altas probabilidades de que una amplia gama de resultados sean posibles, incluyendo resultados muy buenos y resultados tan malos como la extinción humana.[84] Por lo menos, estas muestras sugieren que el tema merece un análisis más detallado.

CAPÍTULO 2

Caminos hacía la superinteligencia

Actualmente las máquinas son muy inferiores a los humanos en inteligencia general. No obstante, algún día (como hemos sugerido) serán superinte- ligentes. ¿Cómo pasaremos de la situación actual a la situación prevista? Este capítulo explorará varios posibles caminos que la tecnología podría tomar. Abordaremos temas como la inteligencia artificial, la estimulación cerebral, la cognición biológica, y los interfaces hombre-máquina, así como las redes y organizaciones. Evaluaremos qué grado de plausibilidad tienen como caminos hacia la superinteligencia. La existencia de múltiples caminos aumenta la probabilidad de que la meta pueda alcanzarse al menos a través de uno de ellos.

Podemos definir tentativamente una superinteligencia como *cualquier intelecto que exceda en gran medida el desempeño cognitivo de los*

humanos en prácticamente todas las áreas de interés[1]. Volveremos sobre el concepto de superinteligencia en el próximo capítulo, donde lo someteremos a una especie de análisis de espectro en el que distinguiremos entre las diferentes formas de superinteligencia posibles. Pero por ahora esta caracterización aproximada será suficiente. Nótese que la definición no dice nada sobre la manera en que la superinteligencia tendrá lugar. Tampoco dice nada sobre sus cualidades: aunque el hecho de que una superinteligencia pueda llegar a tener experiencia consciente subjetiva puede ser muy importante para algunas cuestiones (en particular para cuestiones morales), nuestro interés ahora se centrará en los antecedentes causales y las consecuencias de la superinteligencia, no en teorías metafísicas de la mente[2].

Bajo esta definición, el programa de ajedrez Deep Fritz no es una superinteligen- cia, pues Fritz sólo puede considerarse inteligente en el reducido ámbito del ajedrez. Sin embargo, algunos tipos de inteligencia de ámbito reducido podrían ser relevantes. Cuando nos refiramos a un desempeño superinteligente que se limita a un ámbito particular, remarcaremos explícitamente esa restricción. Por ejemplo, una "superinteligencia ingenieril" sería un intelecto que supera enormemente a las mejores mentes humanas actuales en el campo de la ingeniería. Si no se especifica en otro sentido, utilizaremos el término para referirnos a sistemas que tienen un nivel sobrehumano de inteligencia *general*.

Mas, ¿cómo podríamos llegar a crear superinteligencia? Examinemos algunos posibles caminos.

Inteligencia artificial

Los lectores de este capítulo no deben esperar unas instrucciones detalladas de cómo programar inteligencia artificial general. Dichas instrucciones no existen todavía, por supuesto. Y si estuviera en posesión de tales instrucciones, ciertamente no las publicaría en un libro. (Si las razones para esto no son inmediatamente obvias, los argumentos de capítulos subsiguientes lo aclararán).

Sí podemos, en cualquier caso, discernir algunas características generales del tipo de sistema que sería necesario. A estas alturas parece claro que la capacidad para aprender tendría que ser una característica integral del diseño básico de un sistema pensado para alcanzar la inteligencia general, no algo añadido al final como una extensión u ocurrencia de última hora. Lo mismo ocurre con la habilidad para manejarse frente a la incertidumbre y la información probable. Alguna capacidad para extraer conceptos útiles de los datos sensibles y estados internos, y para utilizar estos conceptos adquiridos dentro de representaciones combinativas usadas en el razonamiento

lógico e intuitivo, también deberían formar parte de las características del diseño básico de una IA moderna pensada para lograr inteligencia general.

Los primeros sistemas de inteligencia artificial a la antigua usanza no se centraron, en su mayoría, en aprender, en manejar incertidumbre o en formar conceptos, quizás porque las técnicas que se ocupan de estas dimensiones estaban poco desarrolladas en aquel momento. Esto no significa que las ideas subyacentes sean completamente nuevas. La idea de usar el aprendizaje como lanzadera capaz de impulsar un sistema simple hasta el nivel de inteligencia humano se remonta al menos hasta el concepto de "máquina infantil" de Alan Turing, sobre la cual escribió en 1950:

> En lugar de tratar de producir un programa que simule la mente adulta, ¿por qué no tratar en su lugar de producir uno que simule la de un niño? Si la sometiéramos entonces a una educación apropiada obtendríamos el cerebro adulto[3].

Turing concibió un proceso repetitivo para desarrollar tal máquina infantil:

> No podemos esperar encontrar una buena máquina infantil al primer intento. Debemos experimentar tratando de enseñar a dicha máquina y ver qué tal aprende. Luego podemos probar con otra y comprobar si es mejor o peor Hay una obvia conexión entre este proceso y la evolución... Podemos esperar, no obstante, que este proceso sea más rápido que la evolución. La supervivencia del más apto es un método lento de medir ventajas. El experimentador, gracias a su inteligencia, debería ser capaz de acelerar el proceso. Igualmente importante es el hecho de que él no estará restringido a mutaciones azarosas. Si puede identificar la causa de alguna debilidad, probablemente podrá idear la mutación que subsanará el problema[4].

Sabemos que procesos evolutivos no-dirigidos pueden producir niveles humanos de inteligencia general, puesto que ya lo han conseguido al menos una vez. Procesos evolutivos dirigidos —esto es, programas genéticos diseñados y guiados por un programador humano inteligente— deberían ser capaces de alcanzar un resultado similar con mucha más eficiencia. Algunos filósofos y científicos, como David Chalmers y Hans Moravec, han utilizado esta observación para defender que la IA de nivel humano no sólo es teóricamente posible sino factible para este siglo[5]. La idea es que podemos estimar las capacidades relativas de la evolución y de la ingeniería humana para producir inteligencia, y llegar a la conclusión de que la ingeniería humana es ya muy superior a la evolución en algunas áreas y que, probablemente, llegue a ser superior en todas las áreas muy pronto. El hecho de que la evolución haya producido inteligencia indica, por lo tanto, que la ingeniería humana llegará pronto a conseguirlo también. Así, Moravec escribió (ya en 1976):

> La existencia de múltiples ejemplos de inteligencia diseñada bajo estas limitaciones

debe darnos mucha confianza en que nosotros podamos conseguir lo mismo en poco tiempo. La situación es análoga a la historia del vuelo en seres más pesados que el aire, en la cual pájaros, murciélagos e insectos demostraron claramente la posibilidad de hacerlo antes de que la cultura lo dominara[6].

Debemos ser precavidos, no obstante, con las inferencias obtenidas mediante estos razonamientos. Es cierto que la evolución produjo vuelo en seres más pesados que el aire, y que ingenieros humanos tuvieron éxito en la misma empresa (si bien por medios de muy diferente naturaleza). Otros ejemplos podrían aducirse, como el sonar, la navegación magnética, las armas químicas, los fotorreceptores, y todo tipo de características mecánicas y cinéticas. Sin embargo, también podríamos fijarnos en las áreas en las que los ingenieros humanos han fracasado igualando a la evolución: en morfogénesis, en auto-reparación y en defensa inmunológica, por ejemplo, los esfuerzos humanos quedan muy lejos de los logros de la naturaleza. El argumento de Moravec, por tanto, no puede darnos "mucha confianza" en que alcanzaremos una inteligencia de nivel humano "muy pronto". En el mejor de los casos, la evolución de vida inteligente se sitúa en un nivel superior dentro de la intrínseca dificultad que conlleva diseñar inteligencia. Pero este nivel superior podría estar bastante por encima de las capacidades actuales de la ingeniería humana.

Otra manera de desplegar un argumento evolutivo en defensa de la posibilidad de la IA es a través de la idea de que podríamos, mediante algoritmos genéticos procesados en ordenadores suficientemente potentes, alcanzar resultados comparables a los de la evolución biológica. Esta versión del argumento evolutivo propone así un método específico mediante el cual la inteligencia podría ser producida.

Pero ¿es cierto que pronto tendremos suficiente potencia computacional para recapitular los procesos evolutivos relevantes que produjeron la inteligencia humana? La respuesta depende de lo que la tecnología computacional avance en las próximas décadas, y de cuánta potencia computacional se necesitará para procesar algoritmos genéticos con la misma capacidad de optimización que el proceso evolutivo de selección natural que reposa en nuestro pasado. Aunque, al final, la conclusión que obtenemos al seguir esta línea de razonamiento es decepcionantemente indeterminado, es productivo lanzar una estimación aproximada (véase cuadro 3). Aunque no consigamos otra cosa, este ejercicio nos permite llamar la atención sobre algunas incógnitas interesantes.

La conclusión es que los recursos computacionales requeridos simplemente para reproducir los procesos evolutivos relevantes que produjeron la inteligencia de nivel humano sobre la tierra están

enormemente lejos de nuestro alcance —y permanecerán así incluso aunque la ley de Moore continuara por un siglo (Cf. Figura 3). Es plausible, no obstante, que comparado con la replicación por fuerza bruta de los procesos de evolución natural, se ganará mucha eficiencia diseñando el proceso de búsqueda que *se dirija* a la inteligencia, empleando varias mejoras obvias frente a la selección natural. Pero es muy difícil estar seguros de la magnitud de esas ganancias

Cuadro 3. *Continúa* 1.200.000 FLOPS. Un modelo multi-compartimental más detallado añadiría entre tres o cuatro órdenes más de magnitud, mientras que modelos de mayor nivel que abstrajeran sistemas de neuronas podrían rebajar entre dos y tres órdenes de magnitud a partir de los modelos más simples[12]. Si quisiéramos simular 10^{25} neuronas a lo largo de más de mil millones de años de evolución (más tiempo del que llevan existiendo los sistemas nerviosos tal y como los conocemos), y dejáramos funcionar a nuestros ordenadores durante un año, tendríamos como resultado un requerimiento de entre $I0^{3I}$-$Í0^{44}$ FLOPS. En comparación, el Tianhe-2 de China, la supercomputadora más potente del mundo a septiembre de 2013, proporciona sólo $3.39xI0^{16}$ FLOPS. En décadas recientes, han sido necesarios aproximadamente 6,7 años para que los ordenadores domésticos aumentaran su potencia en un orden de magnitud. Incluso un siglo de ley Moore continuada no sería suficiente para llenar este vacío. Contar con hardware más especializado, o hacer funcionar los ordenadores más tiempo sólo aumentaría unos pocos órdenes de magnitud[15].

Esta cifra es prudente en otro sentido. La evolución ha alcanzado la inteligencia humana sin proponérsela como objetivo. En otras palabras, las funciones de aptitud para organismos naturales no selecciona sólo en función de la inteligencia y sus precursores[13]. Incluso entornos en los que organismos con habilidades superiores de procesamiento de información se vean recompensados pueden no seleccionar la inteligencia, porque las mejoras en inteligencia pueden (y suelen) imponer costes significativos, tales como un mayor consumo de energía o tiempos más lentos de maduración, y esos costes pueden imponerse a los beneficios obtenidos por un comportamiento más inteligente. Entornos muy hostiles también reducen el valor de la inteligencia: cuanto más corta sea la esperanza de vida, menos tiempo habrá para que se deje notar la recompensa proveniente de una habilidad progresiva para aprender Una reducción de la presión selectiva por la inteligencia hace que se ralentice la expansión de innovaciones mejoradoras de la inteligencia, y así la oportunidad para que la selección favorezca subsiguientes innovaciones que dependen de ella. Y no sólo eso, sino que la evolución podría estancarse en un lugar óptimo que los humanos reconocerían y sobrepasarían mediante la alteración del equilibrio entre explotación y exploración o a través de una progresión de tests de

inteligencia[14] gradualmente más difíciles. Y, como se ha mencionado antes, la evolución disemina mucho de su poder selectivo en rasgos que no están relacionados con la inteligencia (como las carreras de evolución competitiva al estilo Red Queen entre sistemas inmunes y parásitos). La evolución continúa gastando recursos produciendo mutaciones que pueden haberse mostrado consistentemente letales, y fracasa en aprovechar mutaciones distintas con efectos estadísticamente similares. Todo esto son ineficiencias de la selección natural (desde la perspectiva de quien intenta evolucionar la inteligencia) que serían relativamente fáciles de evitar para un ingeniero humano que usara algoritmos desarrolladores de software inteligente.

Es plausible que eliminar ineficiencias como las descritas redujeran varios órdenes de magnitud de los 10^{31}-10^{44} FLOPS calculados anteriormente. Desafortunadamente, es difícil saber exactamente cuántos órdenes de magnitud. Es difícil incluso hacer una estimación aproximada —pues, por lo que sabemos, las ganancias en eficiencia podrían ser de cinco órdenes de magnitud, o diez, o veinticinco.

Figura 3. Rendimiento de las supercomputadoras. En un sentido limitado, “La Ley de Moore” se refiere a la observación de que el número de transistores en circuitos integrados se han doblado durante varias décadas cada dos años aproximadamente. No obstante, el término suele ser usado para referirse a la observación más general de que muchas métricas de desempeño usadas en tecnología computacional han seguido una tendencia exponencial similar Aquí trazamos el pico de velocidad de la supercomputadora más rápida del mundo como una función temporal (en una escala logarítmica vertical). En años recientes, el crecimiento de la velocidad de serie de los procesadores se ha estancado, pero el aumento en el uso de la paralelización ha posibilitado que el número total de computaciones realizadas se mantenga en la tendencia[16].

en eficiencia. No podemos ni siquiera decir si supondrán cinco o veinticinco órdenes de magnitud. En ausencia de ulteriores elaboraciones, los argumentos evolutivos son incapaces de reducir significativamente nuestras expectativas sobre la dificultad de construir inteligencia humana en una máquina, ni pueden hablarnos sobre la escala temporal de tales desarrollos.

Hay otra complicación con este tipo de consideraciones evolutivas, una que hace que sea difícil derivar de ellas siquiera una estimación muy por lo alto de la dificultad de conseguir inteligencia evolutivamente. Debemos evitar el error de inferir del hecho de que vida inteligente evolucionó una vez en la tierra, que los procesos evolutivos implicados tuvieron una probabilidad razonablemente alta de producir inteligencia en un principio. Tal inferencia no es sólida porque no toma en cuenta el efecto de selección observacional que todos los observadores inevitablemente tendrán al haber nacido en un planeta donde surgió vida inteligente, sin importar cómo de probable o improbable era esto para cualquier planeta. Supongamos, por ejemplo, que, junto a los efectos sistemáticos de la selección natural fue necesario una cantidad enorme de *afortunada coincidencia* para

producir vida inteligente —tanto es así que la vida inteligente evoluciona sólo en un planeta de cada 10^3 planetas, donde sólo surgen simples imitaciones. En ese caso, cuando pongamos a funcionar nuestros algoritmos genéticos para replicar lo que la evolución natural hizo, nos podríamos encontrar con que necesitaríamos realizar alrededor de 10^3 simulaciones antes de que encontremos una en la cual todos los elementos se organicen de la manera adecuada. Esto parece totalmente congruente con nuestra apreciación de que la vida sí evolucionó aquí en la Tierra. Sólo mediante razonamientos cuidadosos y algo intrincados —analizando instancias de evolución convergente de rasgos relativos a la inteligencia y enfrentándose a las sutilezas de la teoría de la selección observacional— podemos sortear parcialmente esta barrera epistemológica. A no ser que nos tomemos la molestia de hacer

esto, no estamos en posición de descartar la posibilidad de que la estimación por lo alto de los requerimientos computacionales necesarios para recapitular la evolución de la inteligencia expuesta en el Cuadro 3 sea demasiado baja por treinta órdenes de magnitud (o por otro número igualmente grande[17]).

Otra manera de argumentar a favor de la posibilidad de inteligencia artificial es fijándonos en el cerebro humano y sugiriendo que lo utilicemos como modelo para una inteligencia artificial. Podemos distinguir diferentes versiones de este enfoque basándonos en el grado en que se proponga imitar a las funciones cerebrales biológicas. En un extremo —uno en el que la imitación sería muy cercana— encontramos la idea de *emulación de cerebro completo,* que discutiremos en el siguiente apartado. En el otro extremo encontramos enfoques que se inspiran en el funcionamiento del cerebro pero no buscan una imitación a pequeña escala. Los avances en neurociencia y psicología cognitiva —que se verán ayudados por las mejoras de los instrumentales— deberían eventualmente desentrañar los principios generales del funcionamiento cerebral. Este conocimiento podría entonces guiar los esfuerzos en pos de la IA. Ya hemos mencionado las redes neuronales como un ejemplo de técnica de IA inspirada en el cerebro. La organización perceptual jerarquizada es otra idea trasvasada desde la ciencia cerebral al aprendizaje artificial. El estudio del aprendizaje por refuerzos ha sido motivado (al menos en parte) por su papel en las teorías psicológicas de la cognición animal, y las técnicas de aprendizaje por refuerzo (como por ejemplo el "Algoritmo-TD") inspiradas por estas teorías son ampliamente utilizadas en IA[18]. Con seguridad, más casos como éstos se irán sucediendo a lo largo del tiempo. Puesto que hay un número limitado —quizás un número muy pequeño— de mecanismos fundamentales distintos en el cerebro, un progreso continuo e incremental de la ciencia cerebral debería eventualmente descubrir todos estos mecanismos. Antes de que esto ocurra, sin embargo, es posible que un enfoque híbrido, combinando algunas técnicas inspiradas en el cerebro con algunos métodos puramente artificiales, alcance la meta. Es ese caso, el sistema resultante no tendría por qué ser reconocible como similar al cerebro, incluso aunque algunas claves derivadas del cerebro fueran utilizadas en su desarrollo.

La disponibilidad del cerebro como modelo proporciona un fuerte apoyo a la tesis de que la inteligencia artificial es factible en último término. Esto, no obstante, no nos permite predecir cuándo será alcanzado porque es difícil predecir el ratio de descubrimientos que la ciencia cerebral tendrá en un futuro. Lo que podemos decir es que cuanto más lejos en el futuro miremos, más posible será que veamos los secretos de la funcionalidad cerebral al descubierto hasta el punto

de poder crear inteligencia artificial de esta manera.

Diferentes personas trabajando en pos de la inteligencia artificial mantienen diferentes visiones sobre cómo de prometedoras son las aproximaciones neuromórficas comparadas con aproximaciones que aboguen por diseños completamente sintéticos. La existencia de pájaros demostró que el vuelo de objetos más pesados que el aire era físicamente posible y propició los esfuerzos para construir máquinas voladoras. Pero los primeros aviones funcionales no batían sus alas. No podemos estar seguros de que la inteligencia de las máquinas será como el vuelo, que los humanos consiguieron mediante un mecanismo artificial, o como la combustión, que llegamos a dominar imitando la manera en que ocurren naturalmente los fuegos.

La idea de Turing de diseñar un programa que adquiera la mayoría de su contenido mediante aprendizaje, en lugar de dejarlo pre-programado desde el principio, puede aplicarse igualmente a los enfoques neuromórficos y sintéticos de la inteligencia artificial.

Una variación del concepto de máquina infantil de Turing es la idea de "IA seminal[19]" Mientras que una máquina infantil, como Turing la imaginó, hubiera tenido una arquitectura fija que simplemente desarrollaría sus potencialidades inherentes por la *acumulación de contenido,* una IA seminal sería una inteligencia artificial más sofisticada capaz de mejorar su propia *arquitectura.* En las etapas tempranas de la IA seminal, tales mejoras podrían ocurrir principalmente por ensayo y error, adquisición de información o asistencia por parte de los programadores. En etapas posteriores, no obstante, una IA seminal debería ser capaz de *comprender* su funcionamiento lo suficiente como para producir nuevos algoritmos y estructuras computacionales que impulsaran su desempeño cognitivo. Esta necesaria comprensión podría ser el resultado de que la IA seminal alcanzara un nivel suficiente de inteligencia general en muchos ámbitos, o sobrevenir cuando se superara un cierto umbral en un ámbito particularmente relevante como podría ser la ciencia computacional o las matemáticas.

Esto nos lleva a otro importante concepto, el de "auto-mejoramiento recursivo". Una IA seminal exitosa sería capaz de mejorarse repetidamente a sí misma: una versión temprana de la IA podría diseñar una versión mejorada de sí misma, y la versión mejorada —siendo más inteligente que la original— podría ser capaz de diseñar una versión aún más inteligente de sí misma, y así sucesivamente[20]. Bajo estas circunstancias, tal proceso de auto-mejoramiento recursivo podría continuar lo suficiente como para resultar en una explosión de inteligencia —un evento en el que, en un breve período de tiempo, el nivel de inteligencia de un sistema aumentaría desde una dote modesta de capacidades cognitivas (quizás

sub-humanas en la mayoría de sentidos, pero con un talento específico en codificación y búsqueda de IA) hasta la superinteli- gencia radical. Volveremos sobre esta importante posibilidad en el capítulo 4, donde las dinámicas de tal evento se analizarán pormenorizadamente. Nótese que este modelo indica la posibilidad de sorpresas: los intentos de construir inteligencia artificial general pueden fallar básicamente de plano hasta que el último y crítico factor sea incluido, momento en el cual una IA seminal puede llegar a ser capaz de realizar un auto-mejoramiento recursivo de manera sostenida.

Antes de terminar esta subsección, hay otra cosa que deberíamos enfatizar, y es que una inteligencia artificial no tiene por qué parecerse a una mente humana. IAs podrían ser —de hecho es lo más probable que sean— extremadamente distintas. Deberíamos esperar que tengan arquitecturas cognitivas muy diferentes a las inteligencias biológicas, y en las primeras etapas de su desarrollo tendrán perfiles muy diferentes de fortalezas y debilidades (aunque, como argumentaremos más tarde, éstas podrían sobreponerse a cualquier debilidad inicial). Además, los objetivos sistémicos de las IAs podrían diferir radicalmente de los de los seres humanos. No hay motivo para esperar que una IA genérica esté motivada por el amor o el odio o el orgullo o algún otro típico sentimiento humano: estas complejas adaptaciones requerirían un deliberado y costoso esfuerzo para que se recrearan en las IAs. Esto es a la vez un gran problema y una gran oportunidad. Volveremos sobre la cuestión de la motivación de la IA en capítulos posteriores, pero es tan central al argumento de este libro que merece la pena mantenerlo en mente durante su recorrido.

Emulación de cerebro completo

En la emulación de cerebro completo (también conocida como “subida a la nube”), software inteligente sería producido escaneando y modelando minuciosamente la estructura computacional de un cerebro biológico. Este enfoque representa así un caso límite de inspiración en la naturaleza: el plagio directo. Alcanzar la emulación de cerebro completo requiere seguir los siguientes pasos.

Primero, crear un escaneado del cerebro suficientemente detallado. Esto puede implicar estabilizar el cerebro post-mortem a través de la vidrificación (un proceso que transforma el tejido en una especie de vidrio). Una máquina podría entonces diseccionar el tejido en láminas, que serían entregadas a otra máquina para que las escanearan, quizás por un conjunto de microscopios electrónicos. Diversos colorantes podrían ser aplicados en esta etapa para poner al descubierto diversas propiedades químicas y estructurales. Muchas máquinas escaneadoras podrían trabajar en paralelo para procesar múltiples láminas de cerebro simultáneamente.

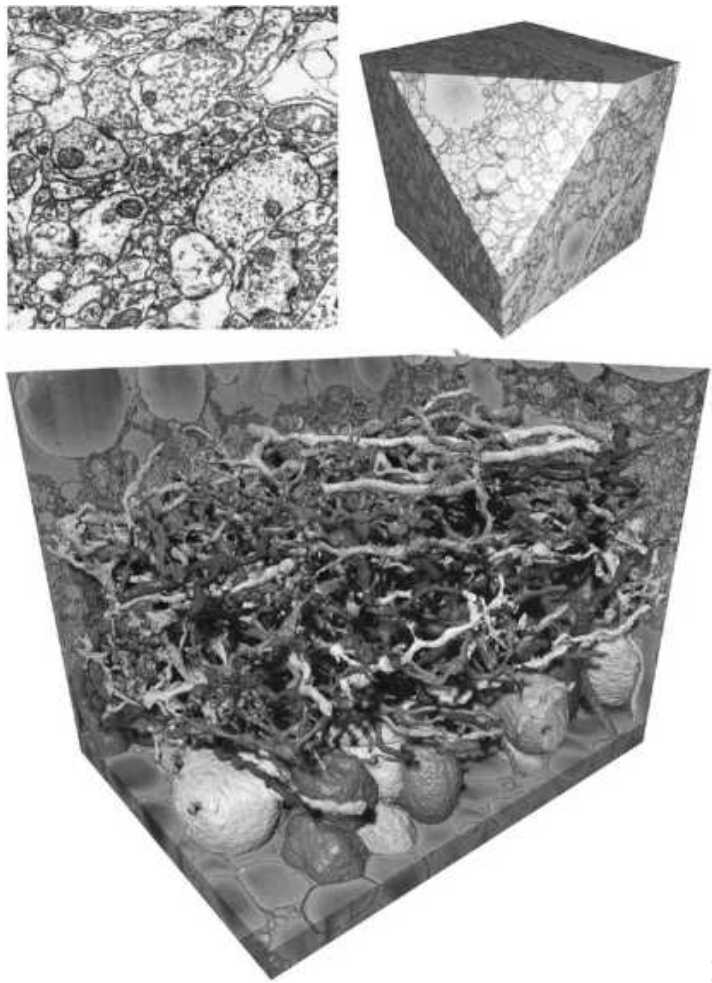
Segundo, los datos en bruto provenientes de los escáneres se pasarían a un ordenador para que realizara un procesamiento automatizado de imágenes que reconstruyera la red neuronal en tres dimensiones que implementó cognición al cerebro original. En la práctica, este paso puede hacerse a la vez que el primero para reducir la cantidad de imágenes en alta resolución almacenadas. El mapa resultante sería entonces combinado con una biblioteca de modelos neurocomputacionales de los diferentes tipos de neuronas o de diferentes elementos neuronales (como tipos particulares de conectores sinápticos). La Figura 4 muestra algunos resultados de los procesos de escaneado y procesamiento de imágenes producidos por tecnología actual.

En la tercera etapa, la estructura neurocomputacional resultante del paso previo es implementada en un ordenador lo suficientemente potente. Si es completamente exitoso, el resultado sería una reproducción digital del intelecto original, con la memoria y la personalidad intacta. La mente humana emulada ahora existiría como software en un ordenador. La mente podría habitar una realidad virtual o interactuar con el mundo exterior mediante apéndices robóticos.

El camino de la emulación de cerebro completo no nos exige que comprendamos cómo funciona la cognición humana ni cómo se programa una inteligencia artificial. Sólo requiere que entendamos características funcionales básicas de los elementos computacionales del cerebro. No es necesario un fundamental avance conceptual ni teórico para el éxito de la emulación de cerebro completo.

La emulación de cerebro completo requiere, sin embargo, de algunas tecnologías realmente avanzadas. Hay tres prerrequisitos clave: (1) *escaneado:* microscopios de alto rendimiento con suficiente resolución y capacidad para detectar propiedades relevantes; (2) *traducción:* análisis automatizado de imágenes que transforme los datos brutos en modelos interpretados tridimensionalmente de los componentes neurona- les relevantes; y (3) *simulación:* hardware suficientemente potente como para imple- mentar la estructura computacional resultante (véase tabla 4). (En comparación con estos difíciles pasos, la construcción de una realidad virtual básica o un cuerpo ro- bótico con un canal audiovisual de recepción y algún canal de salida es relativamente fácil. Una estructura de recepción/salida simple pero mínimamente adecuada parece factible ya con la tecnología del presente[23]).

Figura 4. Reconstrucción 3D neuroana- tómica a partir de imágenes de un microscopio electrónico. *Arriba a la izquierda:* una micrografía electrónica típica que muestra secciones transversales de materias neu- ronales —dendritas y axones—. *Arriba a la derecha:* imagen de volumen de tejido neural de la retina de conejo obtenida por escanea- do block-face de serie mediante un microscopio electrónico.[21]Imágenes individuales en 2D han sido apiladas en un cubo (con un lado de aproximadamente 11 micras). *Conclusión:* Reconstrucción de un subconjunto de las proyecciones neuronales llenando un volumen de neuropilo, generado por un algoritmo de segmentación automatizado[22].



Hay buenas razones para pensar que las tecnologías que harán posible lo anterior son alcanzables, aunque no en un futuro próximo. Ya existen modelos computaciona- les razonables de muchos tipos de neuronas y procesos neuronales. Se ha desarrollado software de reconocimiento de imagen que puede rastrear axones y dendritas a través de una pila de imágenes bidimensionales (aunque es necesario mejorar su fiabilidad). Y hay herramientas de visualización que proporcionan la resolución necesaria —con un microscopio de efecto túnel es posible "ver" los

átomos individuales, en una resolución muy superior a la requerida. Sin embargo, a pesar de que los conocimientos y capacidades actuales parecen sugerir que no existe un impedimento intrínseco al desarrollo de las necesarias tecnologías posibilitantes, está claro que todavía se necesita una gran cantidad de progreso técnico para hacer posible la emulación de cerebro completo humano.[24] Por ejemplo, la tecnología de microscopía necesitaría no sólo suficiente resolución sino también suficiente rendimiento. El uso de un microscopio de escaneado de túnel de resolución atómica para mapear la superficie necesaria sería demasiado lento para ser realizable. Sería más plausible utilizar un microscopio electrónico de menor resolución, pero esto requeriría nuevos métodos para la preparación y tinción de tejido cortical que hicieran visibles detalles relevantes como la estructura sináptica fina. También sería necesaria una gran expansión de las bibliotecas neurocomputacionales y grandes mejoras en el procesamiento automatizado de imágenes e interpretación de escaneado.

En general, la emulación de cerebro completo se basa menos en el conocimiento teórico y más en la capacidad tecnológica que la inteligencia artificial. La cantidad de tecnología necesaria para la emulación cerebro completo depende del nivel de abs-

Tabla 4. Capacidades necesarias para la emulación de cerebro completo

Preparación adecuada de cerebros, cuidadosamente fijados	Preparación adecuada de cerebros, cuidadosamente fijados
Estado físico	Estado físico
Capacidad para escanear volúmenes de imágenes completos en un tiempo y precio razonables	Capacidad para escanear volúmenes de imágenes completos en un tiempo y precio razonables
Escanear a una resolución suficiente como para permitir la reconstrucción	Escanear a una resolución suficiente como para permitir la reconstrucción
Habilidad del escaneado para identificar las propiedades funcionalmente relevantes del tejido	Habilidad del escaneado para identificar las propiedades funcionalmente relevantes del tejido
Ajuste de datos de imperfecciones del escáner	Ajuste de datos de imperfecciones del escáner
Mejora de la calidad del escaneado	Mejora de la calidad del escaneado
Detectar estructuras y procesarla en un modelo 3D consistente del tejido	Detectar estructuras y procesarla en un modelo 3D consistente del tejido
Identificación de cerebros	Identificación de cerebros
Identificación de sinapsis	Identificación de sinapsis
Estimar parámetros funcionalmente relevantes de células, sinapsis y otras entidades	Estimar parámetros funcionalmente relevantes de células, sinapsis y otras entidades
Almacenar el inventario resultante de manera eficiente	Almacenar el inventario resultante de manera eficiente
Modelado de las entidades y su comportamiento	Modelado de las entidades y su comportamiento
Implementación del modelo	Implementación del modelo
Almacenamiento del modelo original y su estado actual	Almacenamiento del modelo original y su estado actual
Comunicación eficiente entre procesadores	Comunicación eficiente entre procesadores
Potencia del procesador para hacer funcionar a la simulación	Potencia del procesador para hacer funcionar a la simulación
Simulación de un cuerpo que posibilite una interacción con un entorno virtual o un entorno real como robot	Simulación de un cuerpo que posibilite una interacción con un entorno virtual o un entorno real como robot
Simulación del entorno	Simulación del entorno

tracción en que se emule el cerebro. En este sentido hay un equilibrio entre conocimiento y tecnología. En general, cuanto peor sea nuestro equipo de exploración y cuanto menos potentes sean nuestros ordenadores, menos podremos confiar en la simulación de procesos electrofisiológicos y químicos del cerebro, y más comprensión teórica de la arquitectura computacional que estamos tratando de emular será necesaria para crear representaciones abstractas de las funcionalidades relevantes.[25] Por el contrario, con una tecnología de escaneo suficientemente avanzada y abundante potencia de cálculo, podría ser posible forzar una emulación incluso con un conocimiento bastante limitado de cerebro. En un irreal caso límite, podríamos imaginar emular un cerebro al nivel de sus partículas elementales utilizando la ecuación de Schrodinger de la mecánica cuántica. En ese caso, se podría confiar enteramente en el conocimiento de la física existente, desechando cualquier modelo biológico. Este caso extremo, sin embargo, demandaría una potencia de cálculo y adquisición de datos imposible. Un nivel mucho más plausible de emulación sería uno que incorporara las neuronas individuales y su matriz de conectividad,

junto con algunas neuronas de la estructura de sus árboles dendríticos y tal vez algunos estados variables de sinapsis individuales. Las moléculas neurotransmisoras no se simularían individualmente, sino que las concentraciones fluctuantes se modelarían de manera poco refinada.

Para evaluar la viabilidad de la emulación de cerebro completo hay que entender cuál es la clave para el éxito. El objetivo no es crear una simulación del cerebro tan detallada y precisa que uno podría utilizarla para predecir exactamente lo que habría ocurrido en el cerebro original si hubiera sido sometido a una secuencia particular de estímulos. El objetivo, en cambio, es captar lo suficiente las propiedades computa- cionalmente funcionales del cerebro para permitir que la emulación resultante pueda llevar a cabo trabajo intelectual. Para este fin, muchos de los caóticos detalles biológicos de un cerebro real son irrelevantes.

Un análisis más elaborado distinguiría entre los diferentes niveles de éxito de emulación basándose en la medida en que se hubiera conservado la funcionalidad de procesamiento de información del cerebro emulado. Por ejemplo, se podría distinguir entre (1) *una emulación de alta fidelidad* que contara con todo el conjunto de conocimientos, habilidades, capacidades y variables del cerebro emulado; (2) *una emulación distorsionada* cuya naturaleza fuera claramente no-humana en algunos aspectos, pero que fuera capaz de hacer el mismo trabajo intelectual que el cerebro emulado; y (3) *una emulación genérica* (que también podría estar distorsionada) que fuera algo así como un niño, que careciera de las habilidades o los recuerdos que habrían sido adquiridas por el cerebro adulto emulado, pero con la capacidad de aprender más de lo que un ser humano normal puede aprender.[26]

Aunque en última instancia parece factible producir una emulación de alta fidelidad, parece bastante probable que la primera emulación de cerebro completo que lograríamos si fuésemos por este camino sería de inferior nivel. Antes de poder conseguir que las cosas funcionaran perfectamente, probablemente se conseguiría que funcionaran de manera imperfecta. También es posible que un impulso hacia la tecnología de emulación diera lugar a la creación de algún tipo de IA neuromórfica que adaptara algunos principios neurocomputacionales descubiertos gracias a los esfuerzos de emulación y así hibridar éstos con métodos sintéticos, siendo posible que esto ocurriera antes de llevar a cabo una emulación de cerebro completo integral y funcional. La posibilidad de que convergiera en una IA neuromórfica de este tipo,

como veremos en un capítulo posterior, hace complicada la evaluación estratégica de la conveniencia de acelerar la tecnología de emulación.

¿Cómo de lejos estamos actualmente de alcanzar una emulación de cerebro completo humano? Una reciente evaluación presentó una hoja de ruta técnica y llegó a la conclusión de que las capacidades previas requeridas podrían estar disponibles a mediados de siglo, aunque con una gran intervalo de incertidumbre.[27] La Figura 5 representa los principales hitos de esta hoja de ruta. Sin embargo, la aparente simplicidad de la hoja de ruta puede ser engañosa, y debemos tener cuidado de no subestimar la cantidad de trabajo que queda por hacer. Ningún cerebro ha sido emulado todavía. Consideremos como modelo el humilde organismo *caenorhab- ditis elegans,* que es un gusano redondo transparente, de aproximadamente 1mm de longitud, con 302 neuronas. La matriz conectiva completa de estas neuronas se conoce desde mediados de la década de 1980, cuando fue laboriosamente trazada por medio del rebanado microscópico electrónico, y de etiquetado a mano de es- pecímenes.[29] Pero saber, simplemente, qué neuronas se conectan con qué otras, no es suficiente. Para crear una emulación del cerebro también tendríamos que saber qué sinapsis son excitadoras y cuáles son inhibidoras; la fuerza de las conexiones; y varias propiedades dinámicas de los axones, las sinapsis, y los árboles dendríticos. Esta información no está disponible todavía ni siquiera para el pequeño sistema nervioso del *c. elegans* (aunque ahora podría estar al alcance de un proyecto de investigación de tamaño moderado).[30] El éxito en la emulación de un cerebro pequeño, como el del *c. elegans*, nos daría una mejor visión de lo que sería necesario para emular cerebros más grandes.

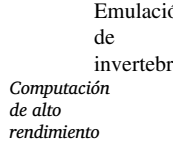


Figura 5. Hoja de ruta de la emulación de cerebro completo. Esquema de aportaciones, actividades y etapas.[28]

En algún momento dentro del proceso de desarrollo de la tecnología, cuando las técnicas llegaran a estar disponibles para emular automáticamente pequeñas cantidades de tejido cerebral, el problema pasará a ser un problema de escala. Nótese "la escalera" en el lado derecho de la Figura 5. Esta serie ascendente de cajas representa una progresión final de cómo los avances pueden dispararse una vez de que los obstáculos preliminares se hayan despejado. Las etapas de esta secuencia se corresponden con las emulaciones de cerebro completo de modelos de organismos neurológicamente más sofisticados, presentados en progresión —por ejemplo, *c. elegans* -> *abeja* -> *ratón* -> macaco *rhesus* -> *humano*. Puesto que las diferencias entre estos peldaños —al menos a partir del primero— son en su mayoría de naturaleza cuantitativa, y debidas principalmente (aunque no del todo) a las diferencias en el tamaño de los cerebros a emularse, éstos deberían ser manejables mediante una ampliación relativamente sencilla de la capacidad de exploración y simulación.[31]

Una vez que empezamos a ascender la escalera final, la eventual consecución de la emulación de cerebro completo humano se vuelve más clara en el horizonte.[32] Podemos, pues, esperar obtener algún aviso previo antes de llegar a la inteligencia artificial de nivel humano si seguimos el camino de emulación del cerebro, al menos si el último requisito en cuanto a tecnologías necesarias para alcanzarla es, o bien la exploración de alto rendimiento, o bien la potencia de cálculo necesaria para la simulación en tiempo real. Sin embargo, si la última tecnología necesaria viene a ser el modelado neurocomputacional, entonces la transición desde prototipos mediocres a una emulación humana funcional podría ser más abrupta. Uno podría imaginar un escenario en el que, a pesar de disponer de abundantes datos de escaneado y potentes ordenadores, resultara difícil conseguir que nuestros modelos neuronales funcionaran bien. Cuando por fin se corte el último fleco, encontraremos que lo que antes era un sistema completamente disfuncional —análogo, tal vez, a un cerebro inconsciente sometido a un grave ataque epiléptico— podría pasar súbitamente a un estado de vigilia coherente. En este caso, el avance decisivo no sería anticipado por una serie de emulaciones animales funcionales de magnitud creciente (provocando titulares de creciente tamaño en los periódicos). Incluso para aquéllos, atentos a estas cuestiones, podría ser difícil saber de antemano cuántos defectos permanecían en los modelos neurocomputacionales previos al éxito en cualquier punto, y cuánto tiempo se iba a tardar en solucionarlos, incluso hasta la víspera del gran avance crítico. (Una vez que se logre una emulación de cerebro completo humano, más desarrollos

potencialmente explosivos se llevarán a cabo, pero hay que aplazar el debate sobre este punto hasta el capítulo 4).

Los escenarios sorpresa son, por tanto, imaginables en el contexto de la emulación de cerebro completo, incluso aunque todas las investigaciones al respecto se lleven a cabo al descubierto. Sin embargo, en comparación con la trayectoria de la IA hacia la inteligencia artificial, la emulación de cerebro completo es más probable que sea prevista claramente, ya que se basa más en tecnologías observables concretas y no está enteramente basada en la especulación teórica. También podemos decir, con mayor confianza que respecto del camino de la IA, que el camino de emulación no tendrá éxito en el futuro cercano (dentro de los próximos quince años, por ejemplo), porque sabemos que varias tecnologías precursoras suponen un gran reto y aún no se han desarrollado. En contraste, parece probable que alguien pudiera, *en principio,* simplemente codificar una IA seminal en una computadora personal de hoy en día; y es concebible —aunque poco probable— que alguien en algún lugar llegue a tener una intuición genial que le permita hacer esto en un futuro cercano.

Cognición biológica

Un tercer camino hacia la inteligencia superior a la inteligencia humana actual es mejorar el funcionamiento de los cerebros biológicos. En principio, esto podría lograrse sin la tecnología, a través de la cría selectiva. Sin embargo, cualquier intento de iniciar un programa clásico de eugenesia a gran escala se enfrentaría a grandes obstáculos políticos y morales. Además, a menos que la selección fuera muy fuerte, se necesitarían muchas generaciones para producir resultados sustanciales. Mucho antes de que tal iniciativa empezara a dar sus frutos, los avances en biotecnología permitirán un control mucho más directo de la genética y neurobiología humanas, dejando a cualquier programa de crianza humana como algo ocioso. Por lo tanto, nos centraremos en los métodos que tienen potencial para ofrecer resultados más rápidos, en una escala temporal de un par de generaciones o menos.

Nuestras capacidades cognitivas individuales pueden fortalecerse de varias maneras, incluyendo métodos tradicionales tales como la educación y la formación. El desarrollo neurológico puede ser promovido a través de intervenciones tecnológicas menores, como la mejora de la nutrición materna e infantil, la eliminación de plomo y otros contaminantes neurotóxicos del medio ambiente, la erradicación de parásitos, el hábito de sueño adecuado y de ejercicio, y la

prevención de enfermedades que afectan al cerebro.[33] Mejoras en la cognición sin duda pueden ser obtenidas a través de cada uno de estos medios, aunque las magnitudes de las ganancias tienden a ser modestas, especialmente en poblaciones que ya están razonablemente bien nutridas y escolarizadas. Desde luego, no lograremos la superinteligencia por cualquiera de estos medios, pero podríamos ayudar marginalmente, sobre todo apoyando a los desfavorecidos y ampliando la captación de talento global. (El declive permanente de la inteligencia debido a la deficiencia de yodo sigue siendo algo generalizado en muchas áreas empobrecidas del interior del mundo — una barbaridad, dado que este problema se puede prevenir proporcionando sal de mesa a los afectados a un coste de unos pocos céntimos por persona y año[34]).

Las mejoras biomédicas podrían propiciar aumentos más grandes. Ya existen fármacos que se suponen que sirven para mejorar la memoria, la concentración y la energía mental en al menos algunos sujetos.[35] (El trabajo que dio lugar a este libro fue impulsado por el café y los chicles de nicotina). Mientras que la eficacia de la presente generación de fármacos inteligentes es variable, marginal, y generalmente dudosa, los nootrópicos futuros podrían ofrecer beneficios más claros y menos efectos secunda- rios.[36] Sin embargo, parece poco plausible, tanto por motivos neurológicos como evolutivos, que se pudiera provocar un aumento dramático de inteligencia introduciendo algunas sustancias químicas en el cerebro de una persona sana.[37] El funcionamiento cognitivo de un cerebro humano depende de la delicada orquestación de muchos factores, especialmente durante las etapas críticas del desarrollo embrionario — y es mucho más probable que esta estructura auto-organizada, para mejorarse, necesite ser cuidadosamente equilibrada, calibrada, y cultivada en lugar de, simplemente, rociarla con alguna poción extraña.

La manipulación genética proporcionará un conjunto de herramientas más potente que la psicofarmacología. Consideremos de nuevo la idea de la selección genética: en lugar de tratar de implementar un programa de eugenesia mediante el control de los patrones de apareamiento, se podría utilizar la selección a nivel de embriones o gametos.[38] El diagnóstico genético pre-implantación ya ha sido utilizado en los procedimientos de fertilización in vitro para la detección de embriones producidos con tendencia a desórdenes monogénicos como la enfermedad de Huntington y con predisposición a algunas enfermedades de desarrollo tardío tales como el cáncer de mama. También se ha utilizado para la selección del sexo y para hacer coincidir el tipo de antígeno leucocitario humano con el de un hermano enfermo, quien entonces se puede beneficiar de una

donación de células madre del cordón umbilical del recién nacido.[39] La gama de rasgos que se pueden seleccionar a favor o en contra se ampliará considerablemente durante la próxima década o dos. Un fuerte motor del progreso en la genética del comportamiento es la creciente caída en el coste de genotipado y secuenciación de genes. El análisis de rasgos genómicos complejos, en estudios con un gran número de sujetos, está empezando a ser factible recientemente, y es algo que aumentará en gran medida nuestro conocimiento de las arquitecturas genéticas de rasgos cognitivos y de comportamiento humano.[40] Cualquier rasgo con una posibilidad no despreciable de ser heredado —incluyendo la capacidad cognitiva— podría entonces ser susceptibles de selección.[41] La selección de embriones no requiere una profunda comprensión de las vías causales por la cual los genes, en la compleja interacción con los entornos, producen fenotipos: requiere sólo (un montón de) datos sobre las correlaciones genéticas de los rasgos de interés.

Es posible calcular algunas estimaciones aproximadas de la magnitud de las ganancias que se pueden obtener en diferentes escenarios de selección.[42] La Tabla 5 muestra incrementos esperables en la inteligencia resultante de diversas formas de selección, todo ello suponiendo que pudiéramos acotar la información completa de las variantes genéticas aditivas comunes que subyacen a la heredabilidad de la inteligencia. (Con información parcial, la eficacia de la selección se reduciría, aunque no del todo hasta el punto que uno podría ingenuamente esperar[44]). De manera poco sorprendente, la selección entre un mayor número de embriones produce mayores ganancias, pero esto conlleva un rendimiento decreciente: la selección entre 100 embriones no produce una ganancia siquiera cerca de cincuenta veces la que se podría obtener de la selección entre 2 embriones.[45]

Tabla 5. Aumentos máximos ***del CI al seleccionar de entre un conjunto de embriones***[43]

Selección	Puntos de CI obtenidos
1 de cada 2	4,2
1 de cada 10	11,5
1 de cada 100	18,8
1 de cada 1000	24,3
5 generaciones (debidos a 1 de cada 10 generaciones)	~65 (debidos a los de 1 de cada 10 generaciones)
Límites acumulativos (variantes aditivas optimizadas para la cognición)	<130 (debido a los rendimientos decrecientes)
	100 + (<300 (debido a los rendimientos decrecientes))

Curiosamente, el rendimiento decreciente se para en gran medida cuando la selección se extiende sobre varias generaciones. Por lo tanto, seleccionar repetidamente al número 1 de 10 durante más de diez generaciones (donde cada nueva generación se compone de los descendientes de los seleccionados en la generación anterior)

producirá un mayor incremento en el valor de rasgo que una selección de 1 de 100 de una sola vez. El problema con la selección secuencial, por supuesto, es que se necesita más tiempo. Si cada paso generacional tarda veinte o treinta años, entonces tan sólo cinco generaciones sucesivas podrían llevarnos hasta bien entrado el siglo XXII. Mucho antes de esto, formas más directas y poderosas de ingeniería genética (por no hablar de inteligencia artificial) estarán probablemente disponibles. Hay, sin embargo, una tecnología complementaria, la cual, una vez que haya sido desarrollada para su uso en seres humanos, potenciaría en gran medida la mejora en la pre-implantación por cribado genético: a saber, la derivación de espermatozoides viables y óvulos a partir de células madre embrionarias.[46] Las técnicas para esto ya se han utilizado para producir descendencia fértil en ratones y células similares a gametos en seres humanos. Sin embargo, todavía persisten algunos retos científicos sustanciales, como la traducción de resultados en animales a seres humanos, o como la prevención de alteraciones epigenéticas en las líneas de células madre derivadas. Según un experto, estos retos pueden tener aplicaciones humanas a “10 o incluso 50 años vista"[47]

Con gametos derivados de células madre, la cantidad de poder selectivo disponible para una pareja podría aumentar en gran medida. En la práctica actual, un procedimiento de fertilización in vitro implica normalmente la creación de menos de diez embriones. Con gametos derivados de células madre, unas pocas células donadas pueden convertirse en un número virtualmente ilimitado de gametos, los cuales pueden ser combinados para producir embriones, que podrían entonces ser genotipados o secuenciados, eligiendo los más prometedores para la implantación. Dependiendo del coste de la preparación y selección de cada embrión individual, esta tecnología podría producir un aumento de múltiple magnitud en el poder selectivo a disposición de las parejas que utilizan la fecundación in vitro.

Más importante aún, los gametos derivados de células madre permitirían, mediante la *selección de embriones por iteración,* comprimir a menos de un período de maduración humana múltiples generaciones de selección. Éste es un procedimiento que constará de los siguientes pasos: [48]

1. Genotipado y selección de un número de embriones destacados en las características genéticas deseadas.
2. Extracción de células madre de esos embriones y conversión de los mismos en espermatozoides y óvulos, madurando dentro de seis meses o menos.[49]
3. Cruzar el nuevo esperma y óvulos para producir embriones.
4. Repetir hasta que se acumulen grandes cambios genéticos.

De esta manera, sería posible llevar a cabo diez o más generaciones de

selección en pocos años. (El procedimiento consumiría mucho tiempo y sería caro; sin embargo, en principio, sólo sería necesario hacerlo una vez en lugar de repetirlo para cada nacimiento. Las líneas celulares conseguidas al final del procedimiento podrían ser utilizadas para generar un gran número de embriones mejorados).

Como indica la Tabla 5, el nivel *medio* de inteligencia entre los individuos concebidos de esta manera podría ser muy alto, posiblemente igual o algo superior a la de la persona más inteligente en la población humana a lo largo de la historia. Un mundo que tuviera una gran población de estos individuos podría (si tuviera la cultura, la educación, la infraestructura de comunicaciones, etc. que lo acompañara) constituir una superinteligencia colectiva.

El impacto de esta tecnología será empañado y retrasado por varios factores. Existe un retraso madurativo inevitable hasta que los embriones finalmente seleccionados se convierten en seres humanos adultos: por lo menos veinte años antes de que un niño mejorado llegue a una productividad total, más todavía antes de que estos niños lleguen a constituir un segmento importante de la población activa. Además, incluso después de que la tecnología se haya perfeccionado, las tasas de adopción probablemente comenzarán a la baja. Algunos países podrían prohibir su uso por completo, en base a criterios morales o religiosos.[50] Incluso allí donde se permita la selección, muchas parejas preferirán la forma natural de concebir. La disposición a utilizar la FIV, sin embargo, podría aumentar si hubiera beneficios claros asociados con el procedimiento —tal como una garantía virtual de que el niño tendría un gran talento y estaría libre de predisposiciones genéticas a la enfermedad—. Menores costes de atención médica y la expectativa de un sueldo vitalicio más alto también servirán de argumento a favor de la selección genética. Si el uso del procedimiento se volviera más común, sobre todo entre las élites sociales, puede haber un cambio cultural en las normas de crianza que hará ver el uso de la selección como algo que las parejas responsables y formadas hacen. Muchos de los inicialmente reacios podrían subirse al carro con el fin de tener un hijo que no está en desventaja respecto de los hijos mejorados de sus amigos y colegas. Algunos países podrían ofrecer incentivos para alentar a sus ciudadanos a tomar ventaja de la selección genética con el fin de aumentar la reserva de capital humano del país, o para aumentar la estabilidad social a largo plazo mediante la selección de rasgos como la docilidad, la obediencia, la sumisión, la conformidad, la aversión al riesgo, o la cobardía, fuera del clan gobernante.

Los efectos sobre la capacidad intelectual también dependerán de la medida en que el poder selectivo disponible sea utilizado para mejorar los rasgos cognitivos (Tabla 6). Los que optan por utilizar

algún tipo de selección de embriones tendrían que elegir la forma de asignar el poder de selección a su disposición, y la inteligencia estaría, en cierta medida, en competencia con otros atributos deseados, como la salud, la belleza, la personalidad, o el atletismo. La selección de embriones por iteración, al ofrecer una gran cantidad de poder selectivo, rebajaría esta competencia, lo que permitiría una fuerte selección simultánea para varios rasgos. Sin embargo, este procedimiento tiende a alterar la relación genética normal entre padres e hijos, algo que podría afectar negativamente a la demanda en muchas culturas.[51]

Con nuevos avances en tecnología genética, puede ser posible sintetizar geno- mas específicos, obviando la necesidad de grandes reservas de embriones. La síntesis de ADN ya es algo rutinario en biotecnología y está en gran medida automatizado, aunque todavía no es factible sintetizar un genoma humano entero que pudiera ser utilizado en un contexto reproductivo (en buena parte debido a las dificultades aún no resueltas para comprender correctamente la epigenética).[54] Pero una vez que esta tecnología haya madurado, un embrión podría ser diseñado con la combinación preferida exacta de rasgos genéticos de cada padre. Los genes que no están presentes en ninguno de los padres también podrían ser empalmados, incluyendo los alelos que

Tabla 6. Posibles impactos de la selección genética en diferentes escenarios[52]

"Fl	[12 puntos]
[más	de 2 embriones [4 puntos]

Sc

impactos directos.

Gr

pr

Ur

au

ha duplicado.

CAMINOS HACIA LA SUPERINTELIGENCIA

están presentes de manera poco frecuente en la población, pero que pueden tener efectos positivos significativos en la cognición.[55]

Una intervención que se hará posible cuando los genomas humanos puedan sintetizarse es la prueba de "corrección ortográfica" genética de un embrión. (La selección de embriones por iteración también podría permitir una versión aproximada de esto). Cada uno de nosotros lleva actualmente una carga mutacional, de cientos de mutaciones quizás, que reducen la eficiencia de los diversos procesos celulares.[56] Cada mutación individual tiene un efecto casi imperceptible (que por ello va lentamente eliminándose del acervo genético); sin embargo, de manera combinada, tales mutaciones puede suponer una gran carga para nuestro funcionamiento.[57] Las diferencias individuales en inteligencia podrían, en una medida significativa, ser atribuibles a las variaciones en el número y naturaleza de tales alelos ligeramente perjudiciales que cada uno de nosotros porta. Con la síntesis de genes podríamos tomar el genoma de un embrión y construir una versión de ese genoma libre de las perturbaciones genéticas provenientes de las mutaciones acumuladas. En términos provocativos, se podría decir que los individuos creados a partir de estos genomas revisados podrían ser "más humanos" que nadie actualmente vivo, en el sentido de que serían expresiones menos distorsionadas de la forma humana. Tales personas no serían puras copias idénticas, porque los seres humanos varían genéticamente de otros modos que no implican a las mutaciones perjudiciales. Pero la manifestación fenotípica de un genoma revisado puede llegar a ser una constitución física y mental excepcional, con un elevado nivel de funcionamiento en dimensiones de rasgo poligénico como la inteligencia, la salud, la resistencia, y la apariencia.[58] (Una vaga analogía podría hacerse con las caras compuestas, en el que los defectos de los individuos superpuestos se equilibran: véase la Figura 6).

Otras técnicas biotecnológicas potenciales también podrían ser relevantes. La clonación humana reproductiva, una vez lograda, se podría utilizar para replicar el ge- noma de individuos excepcionalmente talentosos. La adopción de este sistema estaría limitada por la preferencia de la mayoría de los futuros padres por estar biológica-

Figura 6. Las caras compuestas como una metáfora de corrección ortográfica para genomas. Cada una de las imágenes centrales fue producida por la superposición de fotografías de dieciséis individuos diferentes (residentes de Tel Aviv). Las caras compuestas a menudo se juzgan como más bellas que cualquiera de las caras individuales de que se componen, como si las imperfecciones idiosincrásicas se equilibraran. Análogamente, mediante la eliminación de mutaciones individuales, corregir genomas puede producir gente más cercana a "los ideales platónicos". Tales individuos no serían genéticamente idénticos, ya que muchos genes vienen en múltiples alelos igualmente funcionales. La revisión sólo eliminaría la varianza derivada de mutaciones perjudiciales.[59]

mente relacionados con sus hijos; pero la práctica podría, no obstante, llegar a tener un impacto relevante debido a que (1) incluso un aumento relativamente pequeño en el número de personas excepcionalmente talentosas podrían tener un efecto significativo; y (2) es posible que algún Estado se embarcara en un programa de eugenesia a gran escala, tal vez mediante el pago a madres de alquiler. Otros tipos de ingeniería genética —tales como el diseño de nuevos genes sintéticos o la inserción en el genoma de regiones promotoras y otros elementos de control de la expresión genética— también podrían llegar a ser importantes con el tiempo. Posibilidades aun más exóticas pueden darse, tales como cubas llenas de tejido cortical cultivado de estructura compleja, o animales transgénicos "mejorados" (tal vez algunos mamíferos de gran cerebro, como la ballena o el elefante, enriquecidos con genes humanos). Estos últimos casos son totalmente especulativos, pero en un período de tiempo más largo no pueden ser completamente descontados.

Hasta ahora hemos hablado de las intervenciones en la línea germinal, que se pueden hacer en los gametos o embriones. Mejoras somáticas de genes, sin pasar por el ciclo de la generación, podrían producir, en principio, impacto de manera más rápida. Sin embargo, tecnológicamente son mucho más difíciles. Requieren que los genes modificados puedan insertarse en un gran número de células del cuerpo vivo, incluyendo, en el caso de la mejora de la cognición, en el cerebro. La selección entre las células o embriones de óvulos existentes, en contraste, no requiere la inserción de genes. Incluso este tipo de terapias como la de la línea germinal no implican la modificación del genoma (por ejemplo, la corrección del genoma o el empalme en alelos raros) son mucho más fáciles de implementar en el gameto o la etapa embrionaria, donde se está tratando con un pequeño número de células. Por otra parte, las intervenciones en la línea germinal de embriones probablemente pueden lograr mayores efectos que las intervenciones somáticas en los adultos, debido a que las primeras serían capaces de dar forma al desarrollo temprano del cerebro, mientras que estas últimas estarían limitadas a trabajar sobre una estructura existente. (Algunas de las cosas que podrían hacerse a través de la terapia génica somática también podría ser alcanzables por medios farmacológicos).

Centrándonos, por tanto, en las intervenciones en la línea germinal, debemos tener en cuenta cómo el desfase generacional retrasaría cualquier gran impacto a nivel mundial.[60] Incluso si la tecnología se perfeccionara hoy y empezara a utilizarse de inmediato, se necesitarían más de dos décadas para que una progenie mejorada genéticamente llegara a la madurez. Además, con las aplicaciones en seres humanos normalmente hay una demora de al menos una década

entre la prueba conceptual de laboratorio y la aplicación clínica, debido a la necesidad de extensos estudios que determinen su seguridad. Las formas más simples de selección genética, sin embargo, podrían evitar en gran medida la necesidad de tales pruebas, ya que utilizarían técnicas de tratamiento de fertilidad estándar e información genética para elegir entre embriones que de otro modo habrían sido seleccionados por casualidad.

Los retrasos también pueden ser el resultado de obstáculos no relacionados con un miedo al fracaso (con demandas de pruebas de seguridad), sino relacionados con el miedo al éxito, con la demanda de regulación impulsada por las preocupaciones sobre la legitimidad moral de la selección genética o sus implicaciones sociales más amplias. Tales preocupaciones son probablemente más influyente en algunos países que en otros, debido a los diferentes contextos culturales, históricos y religiosos. La
Alemania de posguerra, por ejemplo, ha optado por dar un gran rodeo frente a las prácticas reproductivas que podrían ser percibidas, incluso en la forma más remota, como dirigidas a la mejora, una postura que es comprensible dada la historia particularmente oscura de atrocidades conectadas al movimiento eugenésico de ese país. Otros países occidentales tienden a adoptar un enfoque más liberal. Y algunos países —tal vez China o Singapur, los cuales tienen políticas de población a largo plazo— podrían no sólo permitir sino promover activamente el uso de la selección genética y la ingeniería genética para mejorar la inteligencia de sus poblaciones, una vez que la tecnología para hacerlo esté disponible.

Cuando el ejemplo se haya establecido, y los resultados empiecen a mostrarse, los reticentes tendrán fuertes incentivos para seguir dicho ejemplo. Las naciones se enfrentarían a la posibilidad de convertirse en remansos cognitivos y quedarse atrás en lo referente a científicos, militares, y concursos de prestigio económicos respecto de los competidores que adoptaran las nuevas tecnologías de mejora humana. Los individuos dentro de una sociedad verían cómo las plazas de las escuelas de élite se llenan de niños seleccionados genéticamente (que también pueden en promedio ser más guapos, más saludables y más conscientes) y querrán tener las mismas ventajas para sus propios hijos. Hay alguna posibilidad de que un gran cambio de actitud aconteciera en un tiempo relativamente corto, tal vez en tan sólo una década, una vez que se haya demostrado que la tecnología funciona y que proporciona un beneficio sustancial. Las encuestas de opinión en Estados Unidos muestran un cambio dramático en la aprobación pública de la fecundación in vitro después del nacimiento del primer “bebé probeta”, Louise Brown, en 1978. Unos años antes, sólo el 18% de los estadounidenses dijo que usarían

en su caso la FIV para tratar la infertilidad; sin embargo, en una encuesta realizada poco después del nacimiento de Louise Brown, el 53% dijo que lo haría, y el número ha seguido subiendo.[61] (En comparación, en una encuesta realizada en 2004, el 28% de los estadounidenses aprueba la selección de embriones para fomentar "la fuerza o la inteligencia", el 58% la aprueba para evitar el cáncer en adultos, y el 68% la aprueba para evitar enfermedades fatales en la infancia.[62])

Si sumamos los diversos retrasos —digamos de cinco a diez años para reunir la información necesaria para que la selección sea significativamente efectiva entre un conjunto de embriones fecundados in vitro (posiblemente mucho más tiempo antes de que los gametos derivados de células estén disponibles para su uso en la reproducción humana), diez años para construir una implantación importante, y de veinte a veinticinco años para que la generación mejorada llegue a la edad en la que comience a ser productiva; nos encontramos con que es poco probable que las mejoras en la línea germinal tengan un impacto significativo en la sociedad antes de mediados de este siglo. Desde ese punto en adelante, no obstante, la inteligencia de segmentos importantes de la población adulta puede comenzar a ser impulsada por mejoras genéticas. La velocidad de este mejoramiento se aceleraría en gran medida cuando grandes grupos de individuos concebidos mediante tecnologías genéticas más potentes y de nueva generación (en gametos derivados de células madre y en particular a través de la selección de embriones por iteración) entren a formar parte de la fuerza laboral.

Con el pleno desarrollo de las tecnologías genéticas descritas anteriormente (dejando a un lado las posibilidades más exóticas como la inteligencia en el tejido neural cultivado), podría ser posible asegurar que los nuevos individuos fueran en promedio más inteligentes que cualquier ser humano que nunca hubiera existido, con posibi-
lidades de llegar más lejos todavía. El potencial de la mejora biológica es, pues, en última instancia, alto, probablemente suficiente para lograr formas leves de super- inteligencia. Esto no debería ser sorprendente. Después de todo, los torpes procesos evolutivos han ampliado dramáticamente la inteligencia en la raza humana, incluso en comparación con nuestros parientes cercanos los grandes simios y con nuestros propios ancestros humanoides; y no hay ninguna razón para suponer que el *homo sapiens* haya alcanzado el vértice de la eficacia cognitiva alcanzable para un sistema biológico. Lejos de ser la especie biológica más inteligente posible, estamos probablemente mejor representados como la especie biológica más estúpida capaz de iniciar una civilización tecnológica —un nicho que llenamos porque

llegamos a ese punto primero, no porque estamos en ningún sentido óptimamente adaptados a ello.

Progresar a lo largo de la ruta biológica es claramente factible. El desfase generacional de las intervenciones en la línea germinal significa que el progreso no puede ser tan repentino y abrupto como en escenarios que impliquen inteligencia artificial. (Las terapias génicas somáticas y las intervenciones farmacológicas podrían saltarse teóricamente el desfase generacional, pero parecen más difícil de perfeccionar y tienen menos probabilidades de producir efectos dramáticos). El potencial último de la inteligencia artificial es, por supuesto, mucho mayor que el de la inteligencia orgánica. (Se puede obtener una idea de la magnitud de la brecha considerando la diferencia de velocidad entre los componentes electrónicos y las células nerviosas: incluso los transistores de hoy en día operan a una velocidad diez millones de veces menor que la de las neuronas biológicas). Sin embargo, incluso las mejoras relativamente moderadas de la cognición biológica podrían tener consecuencias importantes. En particular, la mejora cognitiva podría acelerar la ciencia y la tecnología, incluidos los avances hacia formas más potentes de amplificación de la inteligencia biológica y de la inteligencia artificial. Consideremos cómo el ritmo de avance en el campo de la inteligencia artificial podría cambiar en un mundo donde "hombre promedio Joe" es un par intelectual de igual nivel que Alan Turing o John von Neumann, y donde millones de personas se elevan muy por encima de cualquier gigante intelectual del pasado.[63]

Una discusión sobre las implicaciones estratégicas de la mejora cognitiva tendrá que esperar a un capítulo posterior. Pero podemos resumir esta sección señalando tres conclusiones: (1) por lo menos formas leves de superinteligencia son alcanzables por medio de mejoras biotecnológicas; (2) la viabilidad de humanos cognitivamente mejorados aumenta la probabilidad de que haya formas de inteligencia artificial factibles —porque incluso si *nosotros* fuéramos fundamentalmente incapaces de crear inteligencia artificial (que no hay ninguna razón para suponerlo así), la inteligencia artificial aún podría estar al alcance de humanos cognitivamente mejorados; y (3) si tenemos en cuenta escenarios que se prolongan hasta la segunda mitad de este siglo y más allá, hay que tener en cuenta la aparición probable de una generación de poblaciones genéticamente mejoradas —votantes, inventores, científicos— con la magnitud de la mejora incrementándose rápidamente en décadas posteriores.

Interfaces cerebro-ordenador

En ocasiones se ha propuesto que las interfaces directas de cerebro-ordenador, en particular los implantes, podrían permitir a los seres humanos aprovechar los puntos fuertes de la computación digital —memoria perfecta, cálculo aritmético rápido y preciso, y transmisión de datos por banda ancha— permitiendo al híbrido resultante superar radicalmente al cerebro no aumentado.[64] Pero si bien la posibilidad de conexiones directas entre el cerebro humano y los ordenadores ha sido demostrada, parece poco probable que este tipo de interfaces puedan utilizarse ampliamente como mejoras en ningún momento cercano.[65]

Para empezar, existen riesgos significativos de complicaciones médicas —incluyendo infecciones, desplazamiento del electrodo, hemorragia, y declive cognitivo— cuando se implantan electrodos en el cerebro. Tal vez el ejemplo más vívido hasta la fecha de los beneficios que se pueden obtener a través de la estimulación cerebral es el tratamiento de pacientes con Parkinson. El implante para tratar el Parkinson es relativamente simple: no se comunica realmente con el cerebro sino simplemente suministra una corriente eléctrica estimulante para el núcleo subtalámico. Un vídeo de demostración muestra a un sujeto desplomado en una silla, completamente inmovilizado por la enfermedad, y de repente brota a la vida cuando la corriente se enciende: el sujeto ahora mueve sus brazos, se pone de pie y camina por la habitación, se da la vuelta y realiza una pirueta Sin embargo, incluso detrás de este procedimiento especialmente sencillo y casi milagrosamente exitoso, acechan aspectos negativos. Un estudio de los pacientes con Parkinson que habían recibido implantes cerebrales profundos mostró reducciones en la fluidez verbal, la atención selectiva, el nombramiento de colores, y la memoria verbal en comparación con los sujetos de control. Los sujetos tratados también informaron de otras quejas cognitivas.[66] Tales riesgos y efectos secundarios pueden ser tolerables si el procedimiento se utiliza para aliviar una discapacidad severa. Pero para que los sujetos sanos pidan neurocirugía para sí mismos, tendría que aspirarse a una mejora muy sustancial de la funcionalidad normal.

Esto nos lleva a la segunda razón para dudar de que la superinteligencia se logre mediante la ciborgización, a saber, que es probable que la mejora sea mucho más difícil que la terapia. Los pacientes que sufren de parálisis podrían beneficiarse de un implante que reemplazara sus nervios cortados o que activara patrones generadores de movimiento espinal.[67] Pacientes sordos o ciegos podría beneficiarse de cócleas y retinas artificiales.[68] Pacientes con Parkinson o con dolor crónico podrían beneficiarse de la estimulación cerebral profunda que excita o inhibe la actividad en un área particular del cerebro.[69] Lo que parece mucho más difícil de lograr es una interacción directa de banda ancha entre cerebro y ordenador para

proporcionar aumentos sustanciales en la inteligencia de una forma que no pudiera alcanzarse más fácilmente por otros medios. La mayoría de los potenciales beneficios que los implantes cerebrales podrían proporcionar a sujetos sanos se puede obtener con menos riesgo, gasto y molestia utilizando nuestros órganos motores y sensoriales regulares para interactuar con ordenadores ubicados fuera de nuestros cuerpos. No necesitamos conectar un cable de fibra óptica a nuestro cerebro para acceder a internet. No sólo puede la retina humana transmitir datos a una velocidad impresionante de casi 10 millones de bits por segundo, sino que ésta viene pre-empaquetada con una cantidad masiva de wetware especializado, la corteza visual, que está altamente preparada para extraer significado de ese torrente de información y para interconectarse con otras áreas del cerebro para su posterior procesamiento.[70] Incluso si hubiera una manera fácil de bombear más información en nuestro cerebro, la entrada de datos adicional no haría mucho por aumentar la velocidad a la que pensamos y aprendemos, a no ser que toda la maquinaria neuronal necesaria para dar sentido a los datos se actualizara de manera similar. Dado que esto incluiría a casi todo el cerebro, lo que realmente se necesitaría sería una "prótesis total de cerebro", lo cual es sólo otra forma de decir inteligencia artificial general. Sin embargo, si dispusiéramos de una IA con inteligencia de nivel humano, se podría prescindir de la neurocirugía: una computadora puede tener una carcasa de metal lo mismo que una de huesos. Así que este caso límite sólo nos lleva de vuelta a la senda de la IA, que ya hemos examinado.

La interfaz cerebro-ordenador también se ha propuesto como una manera de sacar información fuera del cerebro, para llevar a cabo comunicaciones con otros cerebros o con máquinas.[71] Tales enlaces han ayudado a los pacientes con síndrome de enclaustramiento a comunicarse con el mundo exterior, permitiéndoles mover un cursor en una pantalla mediante el pensamiento.[72] El ancho de banda alcanzado en tales experimentos es bajo: el paciente minuciosamente escribe una lenta letra tras otra a un ritmo de sólo unas palabras por minuto. Uno puede imaginar fácilmente versiones mejoradas de esta tecnología, tal vez un implante de última generación podría conectarse al área de Broca (una región en el lóbulo frontal que participa en la producción del lenguaje) y reproducir el discurso interno.[73] No obstante, aunque esta tecnología podría ayudar a algunas personas con discapacidades inducidas por accidente cerebrovascular o por degeneración muscular, sería poco atractiva para los sujetos sanos. La funcionalidad que proporcionaría es esencialmente la de un micrófono unido a un software de reconocimiento de voz, que ya está disponible comercialmente —sin el

dolor, molestias, gastos, y los riesgos asociados con la neurocirugía (y sin los implicaciones hiper-orwellianas que supone un dispositivo de escucha intracraneal). Mantener nuestras máquinas fuera de nuestro cuerpo también hace más fácil el actualizarse.

Pero ¿qué pasa con el sueño de sobrepasar el lenguaje por palabras y establecer una conexión entre dos cerebros que permita "descargarse" conceptos de una mente a otra, pensamientos o áreas enteras de experiencia? Podemos descargar archivos de gran tamaño a nuestras computadoras, incluyendo bibliotecas con millones de libros y artículos, y esto se puede hacer en segundos: ¿podría hacerse algo similar con nuestro cerebro? La aparente plausibilidad de esta idea deriva probablemente de una visión incorrecta de cómo se almacena y se representa la información en el cerebro. Como se ha señalado, el obstáculo para alcanzar la inteligencia humana no es la rapidez en que los datos en bruto pueden introducirse en el cerebro, sino más bien lo rápido que el cerebro puede extraer significado y dar sentido a los datos. Tal vez se sugerirá que transmitimos significados directamente, en lugar de paquetes de datos sensoriales que deben ser decodificados por el receptor. Hay dos problemas con esto. La primera es que el cerebro, por contraste con el tipo de programa que normalmente se ejecuta en nuestros ordenadores, no utiliza formatos de almacenamiento y representación de datos estandarizados. Más bien, cada cerebro desarrolla sus propias representaciones idiosincrásicas de contenido de nivel superior. Qué conjuntos neuronales particulares son utilizados para representar un concepto en particular depende de las experiencias únicas del cerebro en cuestión (junto con varios factores genéticos y procesos fisiológicos estocásticos). Al igual que en las redes neuronales artificiales, el significado en redes neuronales biológicas es probable que se represente de manera holística en la estructura y en los patrones de actividad de regiones superpuestas importantes, no en las células separadas de memoria desplegadas claramente.[74] Por lo tanto, no sería posible establecer una correspondencia sencilla entre las neuronas de un cerebro y las de otro, de tal manera que los pensamientos pudieran deslizarse automáticamente de uno al otro. A fin de que los pensamientos de un cerebro sean inteligibles para otro, los pensamientos deben ser descompuestos y empaquetados en símbolos siguiendo alguna convención compartida que permita al cerebro de recepción interpretarlos correctamente. Ésta es la misión del lenguaje.

En principio, podríamos imaginar la descarga del trabajo cognitivo de articulación e interpretación en una interfaz que de algún modo fuera capaz de leer los estados neuronales en el cerebro del emisor y de alguna manera recibir un patrón de medida de activación del cerebro del receptor. Pero esto nos lleva al segundo problema del

escenario ciborg. Incluso dejando de lado el (bastante grande) desafío técnico de cómo leer con fiabilidad y escribir simultáneamente tal vez miles de millones de neuronas de manera individual, crear la interfaz requerida es probablemente un problema de IA completo. La interfaz tendría que incluir un componente capaz de mapear (en tiempo real) patrones de encendido en un cerebro y convertirlos en patrones de encendido semánticamente equivalentes en otro cerebro. La comprensión detallada de múltiples niveles en computación neuronal necesaria para llevar a cabo tal tarea parecería permitir directamente una IA neuromórfica.

A pesar de estas reservas, la ruta ciborg hacia la mejora cognitiva no es del todo sombría. El impresionante trabajo sobre el hipocampo en ratas ha demostrado la viabilidad de una prótesis neural que puede mejorar el rendimiento en la memoria de trabajo sencilla.[75] En su versión actual, el implante recoge las aportaciones de una o dos docenas de electrodos situados en un área ("CA3") del hipocampo y las proyecta en un número similar de neuronas de otro área ("CA1"). Un microprocesador está capacitado para discriminar entre dos patrones de activación diferentes en la primera zona (correspondiente a dos memorias diferentes, "palanca derecha" o "palanca izquierda") y para aprender cómo estos patrones se proyectan en la segunda zona. Esta prótesis no sólo es capaz de restaurar sus funciones cuando se bloquea la conexión neuronal normal entre las dos áreas neuronales, sino que mediante el envío de una señal especialmente clara de un patrón particular de la memoria a la segunda área, se puede mejorar el rendimiento de la memoria más allá de lo que la rata es normalmente capaz de hacer. Aunque sea una gran dificultad técnica para los estándares contemporáneos, el estudio deja muchas preguntas difíciles sin respuesta: ¿Hasta qué punto puede aplicarse este enfoque a un mayor número de memorias? ¿Hasta qué punto podemos controlar la explosión combinatoria, que de no ser controlada amenazaría con hacer que el aprendizaje de asignación correcta fuera imposible a medida que aumentara el número de neuronas de entrada y salida? ¿La mejora de rendimiento en la tarea de prueba trae consigo algún coste oculto, como la disminución de la capacidad de generalizar a partir del estímulo particular usado en el experimento, o la disminución de la capacidad para desaprender la asociación cuando el entorno cambia? ¿Los sujetos de prueba pueden todavía beneficiarse de alguna manera incluso si —a diferencia de las ratas— pudieran acogerse a las ayudas de memoria externas tales como el lápiz y el papel? ¿Y cuánto más difícil sería aplicar un método similar a otras partes del cerebro? Mientras que la presente prótesis se aprovecha de la estructura de alimentación proactiva relativamente simple de las partes del hipocampo (básicamente actúa como un

puente unidireccional entre las áreas CA3 y CA1), otras estructuras de la corteza implican bucles de retroalimentación enrevesados que aumentan en gran medida la complejidad del diagrama de cableado y, presumiblemente, la dificultad de descifrar el funcionamiento de cualquier grupo integrado de neuronas.

Una esperanza para la ruta ciborg es plantear que el cerebro, si se le implantara permanentemente un dispositivo de conexión a algún elemento externo, sería paulatinamente capaz de *aprender* a realizar una asignación eficaz entre sus propios estados cognitivos internos y las entradas o salidas que recibiera del o fueran realizadas a través del dispositivo. En consecuencia, el propio implante no tendría que ser inteligente; más bien, el cerebro se adaptaría inteligentemente a la interfaz, del mismo modo que el cerebro de un niño aprendería gradualmente a interpretar las señales que llegan de los receptores de sus ojos y oídos.[76] Pero aquí de nuevo uno debe preguntarse cuánto realmente se conseguiría. Supongamos que la plasticidad del cerebro fuera tal que podría aprender a detectar patrones de algún nuevo y arbitrario flujo de entrada proyectado sobre una parte de la corteza por medio de una interfaz cerebro-ordenador: ¿por qué no proyectar la misma información sobre la retina en su lugar, como un patrón visual, o sobre la cóclea como sonidos? La alternativa de baja tecnología evita mil complicaciones, y en ambos casos el cerebro podría desplegar sus mecanismos de reconocimiento de patrones y plasticidad para aprender a dar sentido a la información.

Redes y organizaciones

Otro camino concebible hacia la superinteligencia es a través de la mejora gradual de las redes y organizaciones que unen las mentes humanas individuales entre sí y con artefactos y robots. La idea aquí no es que esto mejoraría la capacidad intelectual de las personas lo suficiente como para hacerlos superinteligentes, sino más bien que algún sistema compuesto por individuos en red y organizados podría alcanzar una forma de superinteligencia —lo que en el próximo capítulo vamos a desarrollar bajo el título de "superinteligencia colectiva".[77]

La humanidad ha avanzado enormemente en inteligencia colectiva en el transcurso de la historia y la prehistoria. Los avances provienen de muchas fuentes, incluyendo innovaciones en la tecnología de las comunicaciones, como la escritura y la imprenta; y, sobre todo, de la introducción de la lengua misma; de aumentos en el tamaño de la población mundial y la densidad del asentamiento; de diversas mejoras en las técnicas de organización y normas epistémicas; y de

una acumulación gradual de capital institucional. En términos generales, la inteligencia colectiva de un sistema está limitado por la capacidad de las mentes de sus miembros, por los gastos generales de la comunicación de información relevante entre ellos, y por las diversas distorsiones e ineficiencias que impregnan las organizaciones humanas. Si se reducen los gastos generales de comunicación (incluyendo no sólo los costes de equipo, sino también el retardo en la respuesta, las cargas de tiempo y atención, y otros factores), ulteriores organizaciones más grandes y densamente conectadas se volverían factibles. Lo mismo podría suceder si se encontraran soluciones para algunas de las deformaciones burocráticas que ponen patas arriba la organización de la vida —los juegos de Estado derrochadores, las ampliaciones de presupuesto, la ocultación o falsificación de información y otros problemas administrativos. Incluso las soluciones parciales de estos problemas podrían pagar jugosos dividendos para la inteligencia colectiva.

Las innovaciones tecnológicas e institucionales que podrían contribuir al crecimiento de nuestra inteligencia colectiva son muchas y variadas. Por ejemplo, los mercados de predicción subvencionados podrían fomentar normas de búsqueda de la verdad y mejorar el pronóstico en conflictos científicos y asuntos sociales.[78] Los detectores de mentiras (cuando fuera factible hacer detectores fiables y fáciles de usar) podrían reducir las posibilidades de engaño en los asuntos humanos.[79] Los detectores de autoengaño podrían ser aún más útiles.[80] Incluso sin tecnologías cerebrales novedosas, algunas formas de engaño podría llegar a ser más difícil de practicar gracias a una mayor disponibilidad de muchos tipos de datos, incluyendo el registro de reputación y trayectoria, o la promulgación de normas epistémicas fuertes y de una cultura de racionalidad. La vigilancia voluntaria e involuntaria amasará grandes cantidades de información sobre el comportamiento humano. Las redes sociales ya son utilizadas por más de mil millones de personas para compartir datos personales: en breve, estas personas podrían comenzar a subir las grabaciones continuas de su vida recogidas por los micrófonos y las cámaras de vídeo incorporadas en sus teléfonos inteligentes o en los marcos de sus gafas. El análisis automatizado de dichos flujos de datos permitirá desarrollar muchas aplicaciones nuevas (algunas siniestras y otras benignas, por supuesto).[81]

El crecimiento de la inteligencia colectiva también puede provenir de mejoras organizativas y económicas más generales, y de la ampliación de la fracción de la población mundial que esté educada, que esté digitalmente conectada, y que esté integrada en la cultura intelectual mundial.[82]

Internet se destaca como un horizonte particularmente dinámico

para la innovación y la experimentación. La mayor parte de su potencial puede que todavía permanezca sin explotar. Continuando con el desarrollo de una web inteligente, con un mejor soporte para la deliberación, el desprejuiciamiento, y la agregación de juicios, podrían realizarse grandes contribuciones al aumento de la inteligencia colectiva de la humanidad en su conjunto o en grupos particulares. Pero ¿qué hay de la idea aparentemente más fantasiosa de que internet podría algún día "despertar"? ¿Podría internet convertirse en algo más que la columna vertebral de una superinteligencia colectiva vagamente integrada —algo más parecido a un cráneo virtual que albergaría a una su- perinteligencia emergente unificada? (Esta es una de las formas en que podría surgir la superinteligencia según el influyente ensayo de 1993 de Vernor Vinge, que acuñó el término de "singularidad tecnológica"[83]). Contra esta idea podría objetarse que la inteligencia artificial es bastante difícil de lograr a través de ardua ingeniería, y que no es creíble suponer que vaya a surgir de forma espontánea. Sin embargo, la historia no tiene por qué consistir en que alguna futura versión de internet se convierta de repente en superinteligente de casualidad. Una versión más plausible de este escenario sería que internet acumulara mejoras gracias al trabajo de muchas personas durante muchos años —trabajo en el diseño de mejores algoritmos de búsqueda y filtrado de información, formatos más potentes de representación de datos, agentes de software autónomos más potentes, y protocolos más eficientes que rijan las interacciones entre estos robots— y que la miríada de mejoras incrementales finalmente creara la base de alguna forma más unificada de inteligencia web. Parece por lo menos concebible que un sistema cognitivo basado en la web, sobresaturada con la potencia de los ordenadores y con todos los demás recursos necesarios para un crecimiento explosivo a excepción de un ingrediente fundamental, podría, cuando el componente restante finalmente se arrojara al caldero, ardiera de superinteligencia. Este tipo de escenario, sin embargo, converge en otro posible camino hacia la superinteligencia, el de la inteligencia artificial general, del que ya hemos hablado.

Resumen

El hecho de que haya muchos caminos que conducen a la superinteligencia debería aumentar nuestra confianza en que finalmente llegaremos allí. Si un camino resulta estar bloqueado, aun así podremos avanzar.

Que haya varias rutas no implica que haya múltiples destinos. Incluso si una amplificación de inteligencia significativa se lograra

primero a lo largo de uno de los caminos no mecánicos, esto no dejaría como algo irrelevante a la inteligencia artificial. Todo lo contrario: una mayor inteligencia biológica u organizacional aceleraría los avances científicos y tecnológicos, lo que podría acelerar la llegada de las formas más radicales de amplificación de la inteligencia como la emulación de cerebro completo y la IA.

Esto no quiere decir que sea indiferente cómo se llegue a la superinteligencia artificial. El camino recorrido para llegar allí podría marcar una gran diferencia en el resultado final. Incluso si las capacidades finales que se obtuvieran no dependieran tanto del camino, qué uso se dará a esas capacidades —cuánto control tendrán los seres humanos sobre su disposición— bien podría depender de los detalles de nuestro enfoque. Por ejemplo, las mejoras en inteligencia biológica o en organización podrían aumentar nuestra capacidad para anticipar riesgos y diseñar superinteligencia artificial segura y beneficiosa. (Una evaluación estratégica completa implica muchas complejidades, y tendrá que esperar al capítulo 14).

La verdadera superinteligencia (en contraposición a aumentos marginales en los niveles actuales de inteligencia) podría alcanzarse en primer lugar de manera más plausible a través de la ruta de la IA. Hay, sin embargo, muchas incertidumbres fundamentales a lo largo de este camino. Esto hace que sea difícil evaluar rigurosamente el tiempo que llevará este camino, o cuántos obstáculos hay en el mismo. El camino de la emulación de cerebro completo tiene también alguna posibilidad de ser la ruta más rápida hacia la superinteligencia. Ya que el progreso a lo largo de este camino requiere de avances tecnológicos primariamente incrementales, en lugar de avances teóricos, se puede argumentar con fuerza que este camino tendrá éxito con el tiempo. Parece bastante probable, sin embargo, que incluso si el progreso a lo largo de todo el camino a la emulación cerebral es rápido, la inteligencia artificial, sin embargo, será la primera en cruzar la línea de llegada: esto se debe a la posibilidad de una IA neuromórfica basada en emulaciones parciales.

Las mejoras cognitivas biológicas son claramente factibles, especialmente los basadas en la selección genética. Actualmente, la selección de embriones por iteración parece una tecnología especialmente prometedora. En comparación con los posibles avances en inteligencia artificial, sin embargo, las mejoras biológicas serían relativamente lentas y graduales. Éstas tendrían como resultado, a lo sumo, formas relativamente débiles de superinteligencia (más sobre esto en breve).

La previsible posibilidad de mejora biológica debería aumentar nuestra confianza en que la inteligencia artificial fuera, en última instancia, alcanzable, ya que los científicos e ingenieros humanos

mejorados serán capaces de hacer más progresos y más rápido que sus equivalentes naturales. Especialmente en escenarios en los que la inteligencia artificial se retrase más allá de mediados de siglo, las multitudes cada vez más mejoradas cognitivamente llegadas a escena jugarán un importante papel en la evolución posterior.

Los interfaces cerebro-ordenador parecen una fuente de superinteligencia poco probable. Las mejoras en las redes y organizaciones podrían dar lugar a formas leves de superinteligencia colectiva a largo plazo; pero lo más probable es que jueguen un papel facilitador similar al de la mejora cognitiva biológica, incrementando poco a poco la capacidad efectiva de la humanidad para resolver problemas intelectuales. En comparación con las mejoras biológicas, los avances en las redes y la organización marcarán antes la diferencia —de hecho, estos avances ya se están produciendo continuamente y están teniendo un impacto significativo actualmente. Sin embargo, las mejoras en redes y organizaciones, en comparación con la cognición biológica, sólo pueden proporcionar pequeños aumentos en la capacidad para resolver problemas —aumentando la "inteligencia colectiva" en lugar de la "inteligencia de calidad", anticipando así una distinción que estamos a punto de introducir en el próximo capítulo.

CAPÍTULO 3

Formas de superinteligencia

Entonces, ¿qué entendemos exactamente por "superinteligencia"? **Aunque no deseamos empantanarnos en discusiones terminológicas, hay que decir algo para aclarar el terreno conceptual. Este capítulo identifica tres formas diferentes de superinteligencia y argumenta que son, en un sentido relevante para la práctica, equivalentes. También muestra que hay más posibilidades de que la inteligencia se dé en un sustrato artificial que en un sustrato biológico. Las máquinas tienen una serie de ventajas fundamentales que les darán una superioridad abrumadora. Los humanos biológicos, aun mejorados, serán superados.**

Muchas máquinas y animales no humanos ya se desempeñan a niveles sobrehumanos en ámbitos específicos. Los murciélagos interpretan señales de sónar mejor que el hombre, las calculadoras nos superan en aritmética, y los programas de ajedrez nos ganan al ajedrez. La gama de tareas específicas que un software realiza mejor continuará expandiéndose. Pero si bien los sistemas especializados de procesamiento de información tendrán muchos usos, hay profundas cuestiones adicionales que surgen sólo ante la perspectiva de intelectos artificiales que tuvieran suficiente inteligencia general como para sustituir a los seres humanos en todos los ámbitos.

Como se indicó anteriormente, utilizamos el término "superinteligencia" para referirnos a intelectos que superan considerablemente a las mejores mentes humanas actuales en muchos

dominios cognitivos muy generales. Esto es muy vago todavía. Diferentes tipos de sistemas con atributos de rendimiento muy dispares podrían ser calificados como superinteligencias bajo esta definición. Para avanzar en el análisis, es útil desglosar esta simple noción de superinteligencia distinguiendo diferentes vertientes de supercapacidades intelectuales. Hay muchas maneras para hacer tal descomposición. Aquí vamos a diferenciar entre tres formas: la superinteligencia de velocidad, la superinteligencia colectiva y la superinteligencia de calidad.

Superinteligencia de velocidad

Una superinteligencia de velocidad es un intelecto igual a una mente humana, pero más rápido. Ésta es la forma más fácil de analizar conceptualmente una superinteli- gencia.[1] Podemos definir superinteligencia por velocidad de la siguiente manera:

Superinteligencia de velocidad: *un sistema que puede hacer todo lo que el intelecto humano puede hacer, pero mucho más rápido.*

Por "mucho" queremos decir algo así como "varios órdenes de magnitud" Pero en lugar de tratar de borrar todos los restos de vaguedad de la definición, vamos a confiar en que el lector haga una interpretación razonable.[2] El ejemplo más simple de super- inteligencia de velocidad sería una emulación de cerebro completa que se ejecutara en hardware rápido.[3] Una emulación que funcionara a una velocidad diez mil veces más rápida que un cerebro biológico sería capaz de leer un libro en pocos segundos y escribir una tesis doctoral en una tarde. Con un factor de aceleración de un millón, una emulación podría lograr todo un milenio de trabajo intelectual en un día de trabajo.[4]

Para una mente tan rápida, los acontecimientos en el mundo exterior parecerían desarrollarse a cámara lenta. Suponga que su mente funcionara a 10.000 x. Si un amigo suyo de carne y hueso dejara caer su taza de té, podría ver la porcelana descender lentamente hacia la alfombra en el transcurso de varias horas, como un cometa en silencio deslizándose por el espacio hacia una cita con un planeta lejano; contemplando cómo la anticipación del choque inminente se propaga lentamente a través de los pliegues de la materia gris de su amigo, y de allí hacia su sistema nervioso periférico; podría observar cómo su cuerpo gradualmente toma la forma de un "¡Uy!" congelado —en lo que le daría tiempo no sólo a pedir otra taza, sino también a leer un par de artículos científicos y echarse una siesta.

Debido a esta aparente dilatación del tiempo del mundo material,

una superinte- ligencia de velocidad preferiría trabajar con objetos digitales. Podría vivir en la realidad virtual y hacer frente a dicha economía informacional. Alternativamente, podría interactuar con el medio físico por medio de operadores de nanoescala, ya que las extremidades en escalas tan pequeñas podrían funcionar más rápido que los apéndices macroscópicos. (La frecuencia característica de un sistema tiende a ser inversamente proporcional a su escala de longitud.[5]) Una mente rápida se comunicaría principalmente con otras mentes rápidas en lugar de con tardígrados seres humanos cuasi-vegetales.

La velocidad de la luz se convierte en un obstáculo cada vez más importante cuanto más rápidas se vuelvan las mentes, ya que las mentes más rápidas se enfrentarían a mayores costes de oportunidad a la hora de usar su tiempo para viajar o para comunicarse a larga distancia.[6] La luz es aproximadamente un millón de veces más rápida que un avión a reacción, por lo que se necesitaría un agente digital con una aceleración mental de 1.000.000 x, aproximadamente, la misma cantidad de tiempo subjetivo que un viajero humano contemporáneo necesita para viajar por todo el mundo. Hacer una llamada de larga distancia a alguien tomaría tanto tiempo como llegar allí "en persona", aunque sería más barato porque una llamada requeriría menos ancho de banda. Los agentes con grandes aceleraciones mentales que quisieran conversar extensamente podrían encontrar ventajoso trasladarse unos cerca de otros. Las mentes extremadamente rápidas con necesidad de interacción frecuente (como los miembros de un equipo de trabajo) podrían establecer su residencia en computadoras ubicadas en el mismo edificio para evitar demoras frustrantes.

Superinteligencia colectiva

Otra forma de superinteligencia podría consistir en un sistema que lograra un rendimiento superior mediante la agregación de un gran número de inteligencias menores:

Superinteligencia colectiva: *Un sistema compuesto por un gran número de intelectos menores, de manera que el rendimiento general del sistema superaría enormemente al de cualquier sistema cognitivo actual en muchos ámbitos generales.*

La superinteligencia colectiva es conceptualmente menos clara que la superinteligencia de velocidad.[7] Sin embargo, empíricamente es más familiar. Aunque no tenemos experiencia con mentes de nivel humano que difieran de manera significativa en velocidad, *sí* tenemos amplia experiencia con la inteligencia colectiva, con sistemas compuestos por varios números de componentes de nivel humano que trabajan juntos con diversos grados de eficiencia. Las empresas, los equipos de

trabajo, las redes de cotilleo, los grupos de apoyo, las comunidades académicas, los países, e incluso la humanidad en su conjunto, pueden —si adoptamos una perspectiva un tanto abstracta— ser vagamente definidos como "sistemas" capaces de resolver problemas de tipo intelectual. Por experiencia, tenemos idea de la facilidad con que diferentes tareas acaban siendo imposibles pese a los esfuerzos de organizaciones de distinto tamaño y composición.

La inteligencia colectiva destaca en la resolución de problemas que pueden dividirse fácilmente en partes, de tal manera que las soluciones a los sub-problemas pueden abordarse en paralelo y verificarse independientemente. Tareas como la construcción de un transbordador espacial o el funcionamiento de una franquicia de hamburguesas ofrecen innumerables oportunidades para la división del trabajo: diferentes ingenieros trabajan en los diferentes componentes de la nave; diferentes grupos de personal operan en diferentes restaurantes. En el mundo académico, la rígida división en investigadores, estudiantes, revistas, subvenciones y premios autónomos para distintas disciplinas —aunque poco propicio para el tipo de trabajo que representa este denso libro— podría (sólo en un marco conciliador y amable) ser vista como la base necesaria para encajar los aspectos prácticos que permiten a un gran número de personas y equipos con diversas motivaciones contribuir al crecimiento del conocimiento humano mientras se trabaja con relativa independencia, arando cada uno su propio surco.

La inteligencia colectiva de un sistema puede ser mejorada mediante la ampliación del número o la calidad de sus intelectos constituyentes, o mejorando la calidad de su organización.[8] Para obtener una *superinteligencia* colectiva a partir de cualquier inteligencia colectiva actual se requeriría un grado muy alto de mejora. El sistema resultante tendría que ser capaz de superar enormemente a cualquier inteligencia colectiva de la actualidad o a cualquier sistema cognitivo similar en muchos ámbitos muy generales. Un nuevo formato de conferencia que permitiera a los investigadores intercambiar información de manera más efectiva, o un nuevo algoritmo de colaboración de filtrado de información que predijera mejor las calificaciones de los usuarios de libros y películas, claramente no podría por sí solo llevarnos a algo parecido a una superinteligencia colectiva. Tampoco lo haría un aumento del 50% en la población mundial, o una mejora en el método pedagógico que permitiera a los estudiantes completar un día de clases en cuatro horas en lugar de seis. Se necesitaría un crecimiento mucho más extremo de la capacidad cognitiva colectiva de la humanidad para cumplir con el criterio de superinteligencia colectiva.

Nótese que el umbral para la superinteligencia colectiva está

relacionado con los niveles de rendimiento presentes, es decir, de principios del siglo XXI. A lo largo de la prehistoria humana, y de nuevo durante el curso de la historia humana, la inteligencia colectiva de la humanidad ha crecido en factores muy grandes. La población mundial, por ejemplo, ha aumentado por lo menos en un factor de mil desde el Pleis- toceno.[9] Sobre esta base, los niveles actuales de inteligencia colectiva humana podrían ser considerados cercanos a la superinteligencia en relación a la *línea de base del Pleis- toceno.* Algunas mejoras en comunicaciones tecnológicas —especialmente el idioma hablado, pero tal vez también las ciudades, la escritura y la impresión—, podrían asimismo ser consideradas como precursoras, en su forma individual o en combinación, de impulsos muy grandes, en el sentido de que si otra innovación de impacto comparable a nuestra capacidad colectiva de resolución de problemas intelectuales llegara a suceder, se traduciría en la superinteligencia colectiva.[10]

Algunos lectores se verán tentados en este punto de argumentar que la sociedad moderna no parece tan inteligente. Tal vez alguna decisión política indeseada se haya producido en el país de origen del lector, y la aparente falta de sentido común de esa decisión ocupa ahora un lugar preponderante en la mente del lector como evidencia de la incapacidad mental de la era moderna. ¿Y no es verdad que la humanidad contemporánea está idolatrando el consumo de materiales, agotando los recursos naturales, contaminando el medio ambiente, diezmando a la diversidad de especies, y todo mientras falla en remediar evidentes injusticias mundiales y dejando de lado los valores humanísticos o espirituales más básicos? Sin embargo, dejando de lado la cuestión de cómo se comparan las deficiencias de la modernidad frente a las deficiencias no tan despreciables de épocas anteriores, no hay nada en nuestra definición de superinteligencia colectiva que implique que una sociedad con una mayor inteligencia colectiva sea necesariamente mejor. La definición ni siquiera implica que la sociedad más inteligente de manera colectiva sea más *sabia.* Podemos pensar en la sabiduría como la capacidad de hacer que las cosas importantes salgan de manera más o menos correcta. Entonces es posible imaginar una organización compuesta por un gran grupo de trabajadores del conocimiento coordinados de manera muy eficiente, que en conjunto pudieran resolver problemas intelectuales en muchos ámbitos muy generales. Esta organización, supongamos, podría funcionar en la mayoría de tipos de empresa, inventar la mayoría de tipos de tecnologías, y optimizar la mayoría de tipos de proceso. Aun así, podríamos sacar algunas conclusiones completamente erróneas sobre cuestiones clave de perspectiva global —por ejemplo, podríamos dejar de tomar las precauciones adecuadas contra los riesgos existenciales— y como resultado llevar a cabo un corto período de

crecimiento explosivo que terminara sin pena ni gloria en un colapso total. Tal organización podría tener un alto grado de inteligencia colectiva; si es lo suficientemente alta, la organización sería una superinteligencia colectiva. Debemos resistir la tentación de subsumir todos los atributos normativamente deseables en un enorme concepto amorfo de funcionamiento mental, como si uno nunca pudiera considerar admirable un rasgo sin que los demás estuvieran igualmente presentes. En su lugar, debemos reconocer que pueden existir sistemas de procesamiento de información instrumentalmente poderosos —sistemas inteligentes— que no son ni intrínsecamente buenos ni sabios de manera fiable. En todo caso, examinaremos esta cuestión en el capítulo 7.

La superinteligencia colectiva podría estar integrada de manera vaga o concreta. Para ilustrar un caso de superinteligencia colectiva vagamente integrada, imaginemos un planeta, *Megatierra,* que tiene el mismo nivel de tecnologías de comunicación y coordinación que tenemos actualmente en la Tierra real, pero con una población un millón de veces más grande. Con una población tan enorme, la fuerza de trabajo intelectual total en Megatierra sería proporcionalmente mayor que en nuestro planeta. Supongamos que un genio científico del calibre de Newton o Einstein surge al menos una vez por cada 10 millones de personas: entonces en Megatierra habría 700.000 de esos genios que vivirían contemporáneamente, junto a vastas multitudes de talentos levemente menores. Las nuevas ideas y tecnologías se desarrollarían a un ritmo vertiginoso, y la civilización mundial de Megatierra constituiría una superinteligencia colectiva vagamente integrada.[11]

Si aumentamos gradualmente el nivel de integración de una inteligencia colectiva, podría llegar a convertirse en un *intelecto* unificado —una sola y enorme "mente" en lugar de un simple conjunto de mentes humanas más pequeñas interactuando libremente.[12] Los habitantes de Megatierra podrían dar pasos en esa dirección mediante la mejora de las comunicaciones y las tecnologías de coordinación y mediante el desarrollo de mejores formas de trabajo para muchas personas que trabajan conjuntamente en cualquier problema intelectual. Una superinteligencia colectiva podría, por lo tanto, tras ganar lo suficiente mediante la integración, convertirse en una "superin- teligencia de calidad."

Superinteligencia de calidad

Podemos distinguir una tercera forma de superinteligencia.

Superinteligencia de calidad: *un sistema que es al menos tan rápido como una mente*

humana y cualitativamente mucho más inteligente.

Al igual que la inteligencia colectiva, la calidad de la inteligencia es también un concepto un tanto turbio; y en este caso la dificultad se ve agravada por nuestra falta de experiencia con alguna variación de calidad en inteligencia por encima del extremo superior de la distribución humana actual. Podemos, sin embargo, conseguir un poco de comprensión sobre esta noción considerando algunos casos relacionados.

En primer lugar, podemos ampliar la gama de nuestros puntos de referencia teniendo en cuenta a los animales no humanos, que tienen una inteligencia de menor calidad. (Esto no implica una concepción especista. Un pez cebra tiene una calidad de inteligencia que se adapta de manera excelente a sus necesidades ecológicas; pero la perspectiva relevante aquí es más antropocéntrica: nuestra preocupación tiene que ver con el rendimiento en tareas cognitivas complejas *humanamente* relevantes). Los animales no humanos carecen de lenguaje estructurado complejo; no son capaces de utilizar, o de utilizar sólo rudimentariamente, herramientas, ni de construir herramientas; están severamente restringidos en su capacidad de hacer planes a largo plazo; y tienen una capacidad de razonamiento abstracto muy limitada. Estas limitaciones tampoco se explican completamente por la falta de inteligencia de velocidad o de inteligencia colectiva entre las mentes de los animales no humanos. En términos de potencia de cálculo bruto, los cerebros humanos son probablemente inferiores a los de algunos animales grandes, incluyendo los elefantes y las ballenas. Y aunque la compleja civilización tecnológica de la humanidad no sería posible sin nuestra enorme ventaja en inteligencia colectiva, no todas las capacidades cognitivas claramente humanas dependen de la inteligencia colectiva. Muchas están ya muy desarrolladas en

hordas pequeñas y aisladas de cazadores-recolectores.[13] Y muchas no están tan desarrolladas entre los animales no humanos altamente organizados, como los chimpancés y los delfines intensamente entrenados por instructores humanos, o las hormigas que viven en sus grandes y bien ordenadas sociedades. Evidentemente, los notables logros intelectuales del *homo sapiens* son en gran medida atribuibles a las características específicas de nuestra arquitectura cerebral, características que dependen de una dotación genética única, no compartida por otros animales. Esta observación puede ayudar a ilustrar el concepto de superinteligencia de calidad: sólo estamos ante una superinteligencia de calidad si es al menos tan superior a la inteligencia humana en cuanto a calidad como la inteligencia humana lo es respecto de la de los elefantes, delfines, o chimpancés.

Una segunda manera de ilustrar el concepto de superinteligencia de calidad es observando los déficits cognitivos en ámbitos específicos que pueden afectar a los seres humanos individuales, en particular los déficits que no son causados por demencia general o por otras condiciones asociadas con la destrucción total de los recursos neurocomputacionales del cerebro. Consideremos, por ejemplo, a las personas con trastornos del espectro autista que pueden tener déficits notables en la cognición social, mientras que se desempeñan bien en otros dominios cognitivos; o personas con amusia congénita, que no pueden tararear o reconocer melodías simples aunque funcionen con normalidad en la mayoría de los demás aspectos. Muchos otros ejemplos podrían aducirse de la literatura neuropsiquiátrica, que está repleta de casos de estudio de pacientes que sufren déficits estrictamente circunscritos causados por anomalías genéticas o por traumas cerebrales. Estos ejemplos muestran que los adultos humanos normales tienen una gama de talentos cognitivos notables que no se tienen simplemente en función de poseer una cantidad suficiente de potencia de procesamiento neural general o incluso en función de poseer una cantidad suficiente de inteligencia general: también se necesitan circuitos neuronales especializados. Esta observación sugiere la idea de que existan *talentos cognitivos posibles, pero no desarrollados,* que ningún ser humano real posee a pesar de que otros sistemas inteligentes —sistemas sin más potencia de procesamiento que el cerebro humano— que sí tuvieran esos talentos se beneficiarían enormemente de la capacidad para llevar a cabo una amplia gama de tareas estratégicamente relevantes.

En consecuencia, teniendo en cuenta a los animales no humanos y a los individuos humanos con déficits cognitivos en ámbitos específicos, podemos formarnos una idea de las diferentes calidades de inteligencia y de la diferencia práctica que producen. Si el *homo sapiens* hubiera carecido (por ejemplo) de los módulos cogniti- vos que

permiten representaciones lingüísticas complejas, podría haber sido sólo otra especie de simio que viviría en armonía con la naturaleza. Por el contrario, si llegáramos a conseguir un nuevo conjunto de módulos que nos dieran una ventaja similar a la de ser capaz de formar representaciones lingüísticas complejas, llegaríamos a ser superinteligentes.

Alcance directo e indirecto

Una superinteligencia en cualquiera de estas formas podría, con el tiempo, desarrollar la tecnología necesaria para crear cualquiera de las otras. El *alcance indirecto* de estas

tres formas de superinteligencia es, por tanto, el mismo. En este sentido, el alcance indirecto de la inteligencia humana actual es similar, si suponemos que finalmente seremos capaces de crear algún tipo de superinteligencia. Sin embargo, hay un sentido en el que las tres formas de superinteligencia están mucho más cerca una de la otra: cualquiera de ellas podría crear otras formas de superinteligencia más rápidamente de lo que nosotros podríamos hacer desde nuestro estado actual.

Los *alcances directos* de las tres formas diferentes de superinteligencia son más difíciles de comparar. Puede que no haya un orden definitivo. Sus respectivas capacidades dependen del grado en que se plasmen sus respectivas ventajas —*cómo* de rápida sea una superinteligencia de velocidad, *cómo* de cualitativamente superior sea una superinteligencia de calidad, y así sucesivamente. A lo sumo, podríamos decir que, *ceteris paribus,* la superinteligencia de velocidad destaca en tareas que requieren la rápida ejecución de una larga serie de pasos que deben realizarse secuencialmente, mientras que la superinteligencia colectiva destaca en tareas que permiten la descomposición analítica en sub-tareas paralelas y en tareas que exigen la combinación de diferentes perspectivas y habilidades. En algún sentido vago, la superinteligencia de calidad sería la forma más capaz de todas, en la medida en que podría comprender y resolver problemas que están, en un sentido práctico, fuera del alcance *directo* de la superinteligencia de velocidad y de la superinteligencia colectiva.[14]

En algunos ámbitos, la cantidad es un mal sustituto de la calidad. Un genio solitario trabajando en una habitación descorchada puede escribir *En busca del tiempo perdido*. ¿Podría una obra maestra equivalente producirse mediante la contratación de un edificio de oficinas lleno de escritorzuelos?[15] Incluso dentro de la gama actual de variación humana vemos que algunas funciones se benefician enormemente de la mano de obra de un solo autor intelectual brillante en comparación a los esfuerzos conjuntos de muchos mediocres. Si ampliamos nuestro espectro para incluir en él mentes superinteligentes, debemos contemplar la posibilidad de que haya problemas intelectuales que sólo sean solucionables por una superinteligencia y que sean inasequibles para cualquier conjunto, del tamaño que sea, de seres humanos no mejorados.

Puede, por lo tanto, haber algunos problemas que fueran solucionables por una superinteligencia de calidad, y tal vez por una superinteligencia de velocidad, pero que una superinteligencia colectiva poco integrada no pudiera resolver (a no ser que primero amplificara su propia inteligencia).[16] No podemos ver claramente la naturaleza de todos estos problemas, pero podemos caracterizarlos en

términos generales.[17] Éstos tienden a ser problemas que implican múltiples interdependencias complejas que no permiten soluciones independientemente verificables: problemas que, por lo tanto, no pueden ser resueltos de una manera gradual, y para los que podríamos requerir nuevas clases de comprensión cualitativa o nuevos marcos de representación que son demasiado profundos o demasiado complicados para que el conjunto de los actuales mortales los descubran o los usen de manera efectiva. Algunos tipos de creación artística y de cognición estratégica podrían encajar en esta categoría. También algunos tipos de avance científico, tal vez. Y uno puede especular que la lentitud e inseguridad del progreso de la humanidad en muchos de los "problemas eternos" de la filosofía se deben a la falta de adaptación de la corteza humana para el trabajo filosófico. En este punto de vista, nuestros filósofos más célebres son como los perros que caminan sobre sus patas traseras —sólo alcanzan por poco el umbral de rendimiento requerido *simplemente* para llevar a cabo el ejercicio de la actividad.[18]

Fuentes de ventaja para la inteligencia digital

Pequeños cambios en el volumen cerebral y en el cableado pueden tener importantes consecuencias, como vemos cuando comparamos los logros intelectuales y tecnológicos de los seres humanos con los de otros simios. Los cambios mucho más grandes en recursos informáticos y arquitectura que la inteligencia artificial permitirá, probablemente tendrán consecuencias aún más profundas. Es difícil para nosotros, quizás imposible, imaginarnos en un sentido intuitivo las aptitudes de una superinteligencia; pero al menos podemos tener una idea del rango de sus posibilidades examinando algunas de las ventajas que tendrían las mentes digitales. Las ventajas de hardware son más fáciles de apreciar:

- *Velocidad de los elementos computacionales.* Las neuronas biológicas operan a una velocidad máxima de aproximadamente 200 Hz, un total de siete órdenes de magnitud más lento que un microprocesador moderno (~ 2 GHz).[19] Como consecuencia de ello, el cerebro humano se ve obligado a confiar en la paralelización masiva y es incapaz de llevar a cabo rápidamente cualquier cálculo que requiera un gran número de operaciones secuenciales.[20] (Cualquier cosa que el cerebro hace en menos de un segundo no puede requerir más de un centenar de operaciones secuenciales, quizás sólo unas pocas docenas). Sin embargo, muchos de los algoritmos más importantes en sentido práctico para la programación y para la informática no pueden paralelizarse de manera fácil. Muchas de las tareas cognitivas podrían realizarse mucho más eficientemente si el soporte nativo del cerebro para algoritmos paralelizables buscadores de patrones se complementara con, y estuviera integrado por, ayudas para el procesamiento secuencial rápido.
- *Velocidad de comunicación interna.* Los axones llevan potenciales de acción a una velocidad de 120 m/s o menos, mientras que los núcleos de

procesamiento electrónicos pueden comunicarse de manera óptica a la velocidad de la luz (300.000.000 m/s).[21] La lentitud de las señales neuronales limita cómo de grande puede ser un cerebro biológico al mismo tiempo que funciona como una unidad de procesamiento individual. Por ejemplo, para conseguir un retardo de ida y vuelta de menos de 10 ms entre dos elementos de un sistema, los cerebros biológicos deben ser menores que 0,11 m^3. Un sistema electrónico, por otro lado, podría ser de 6,1 x $I0^{17}$ m^3, aproximadamente del tamaño de un planeta enano: dieciocho órdenes de magnitud más grande.[22]

- *Número de elementos computacionales.* El cerebro humano tiene un poco menos de 100 mil millones de neuronas.[23] Los seres humanos tienen un cerebro casi tres veces y media más grande que el de los chimpancés (aunque sólo una quinta parte del tamaño del cerebro de los cachalotes).[24] El número de neuronas en una criatura biológica está obviamente limitado por su volumen craneal y sus limitaciones metabólicas, pero también por otros factores que pueden ser significativos para cerebros más grandes (como la refrigeración, el tiempo de desarrollo, o la demora en la señal conductiva —véase el punto anterior). Por el contrario, el hardware es indefinidamente escalable hasta límites físicos muy altos.[25] Las supercomputa- doras pueden tener el tamaño de un almacén o más, con capacidad adicional remota añadida a través de cables de alta velocidad.[26]
- *La capacidad de almacenamiento.* La memoria de trabajo humano es capaz de mantener no más de cuatro o cinco bloques de información en cualquier momento dado.[27] Aunque sería erróneo comparar el tamaño de la memoria de trabajo humana directamente con la cantidad de RAM de una computadora digital, es evidente que las ventajas de hardware de las inteligencias digitales harían posible que éstas tuvieran memorias de trabajo más grandes. Esto podría permitirle a esas mentes entender intuitivamente relaciones complejas que los seres humanos sólo pueden manejar tentativamente a través de calculaciones perseverantes.[28] La memoria humana a largo plazo también es limitada, aunque no está claro si logramos agotar su capacidad de almacenamiento durante el curso de una vida ordinaria —la velocidad a la que vamos acumulando información es muy lenta. (En una estimación, los adultos humanos guardan en el cerebro unos mil millones de bits— un par de órdenes de magnitud menos que un smartphone de gama baja[29]). Tanto la cantidad de información almacenada como la velocidad con la que se puede acceder sería, por tanto, mucho mayor en un cerebro artificial que en un cerebro biológico.
- *Fiabilidad, vida útil, sensores, etc.* Las inteligencias artificiales pueden tener varias otras ventajas de hardware. Por ejemplo, las neuronas biológicas son menos fiables que los transistores.[30] Mientras que la compleja computación requiere esquemas de codificación redundantes que utilizan múltiples elementos para codificar un solo bit de información, un cerebro digital podría conseguir algunas mejoras a través de la eficiencia en la utilización de elementos de computación fiables y de alta precisión. Los cerebros se fatigan después de unas horas de trabajo y comienzan a decaer de forma permanente después de unas décadas de tiempo subjetivo; los microprocesadores no están sujetos a estas limitaciones. El flujo de datos en una inteligencia artificial podría aumentarse mediante la adición de millones de sensores. Dependiendo de la tecnología utilizada, una máquina podría tener hardware re- configurable que podría ser optimizado para cambiar en función de los requisitos de la tarea, mientras que gran parte de la arquitectura del cerebro está fijada desde el

nacimiento o se cambia lentamente (aunque algunos detalles de la conectividad sináptica pueden cambiar en escalas de tiempo más cortas, como días).[31]

En la actualidad, el poder computacional del cerebro biológico todavía se compara favorablemente con el de las computadoras digitales, aunque la supercomputadoras punteras están alcanzando niveles de rendimiento que están dentro del rango de estimaciones plausibles de potencia de procesamiento del cerebro.[32] Pero el hardware está mejorando rápidamente, y los límites últimos de rendimiento para el hardware son muy superiores a los de los sustratos biológicos de computación.

Las mentes digitales también se beneficiarán de importantes ventajas de software:

- *Editabilidad.* Es más fácil experimentar con variaciones de parámetros en software que en wetware neural. Por ejemplo, con una emulación de cerebro completo uno podría fácilmente juzgar qué sucede si se añaden más neuronas a un área cortical en particular o si se aumenta o disminuye la excitabilidad. La ejecución de experimentos similares con cerebros biológicos vivos sería mucho más difícil.
- *Capacidad de duplicación.* Con software, uno puede hacer rápidamente y de manera arbitraria muchas copias de alta fidelidad para llenar la base de hardware disponible. Los cerebros biológicos, por el contrario, sólo pueden reproducirse muy lentamente; y cada nueva instancia comienza en un estado de indefensión, sin recordar nada de lo que sus padres aprendieron en sus vidas.
- *Coordinación hacia la meta.* Los colectivos humanos están repletos de ineficiencias derivadas del hecho de que es casi imposible conseguir una uniformidad completa de propósito entre los miembros de un grupo grande, por lo menos hasta que se vuelva factible Inducir la docilidad a gran escala mediante drogas o selección genética. Un "clan de copias" (un grupo de programas idénticos o casi idénticos que comparten un objetivo común) podría evitar este tipo de problemas de coordinación.
- *Compartir memoria.* Los cerebros biológicos necesitan largos períodos de formación y tutoría mientras que las mentes digitales podrían adquirir nuevos recuerdos y habilidades mediante el canje de archivos de datos. Una población de mil millones de copias de un programa de IA podría sincronizar sus bases de datos periódicamente, de modo que todas las copias del programa sabrían todo lo que cualquier copia aprendió durante la hora anterior (La transferencia directa de memoria requiere formatos de representación normalizados. Un intercambio fácil de contenido cognitivo de alto nivel, por lo tanto, no sería posible entre cualquier par de inteligencias artificiales. En particular, no sería posible entre la primera generación de emulaciones de cerebro completo).
- *Nuevos módulos, modalidades, y algoritmos.* La percepción visual nos parece algo fácil y sin esfuerzo, muy diferente a resolver problemas de geometría, a pesar de que se necesita una cantidad masiva de computación para reconstruir, a partir de los patrones bidimensionales de estimulación de nuestras retinas, una representación tridimensional de un mundo poblado de objetos reconocibles. La razón de que esto nos parezca fácil es que hemos dedicado maquinaria neural de pequeña escala para el

procesamiento de información visual. Este procesamiento a pequeña escala se produce inconscientemente y de forma automática, sin necesidad de agotar nuestra energía mental o nuestra atención consciente. La percepción de la música, el uso del lenguaje, la cognición social, y otras formas de procesamiento de información que son "naturales" para nosotros los seres humanos, parecen estar igualmente apoyadas en módulos neurocomputacionales específicos. Una mente artificial que tuviera ese apoyo especializado para otros ámbitos cognitivos que han llegado a ser importantes en el mundo contemporáneo —como la ingeniería, la programación de ordenadores, y la estrategia empresarial— tendría grandes ventajas sobre mentes como las nuestras que dependen de una torpe cognición de propósito general para pensar esas cosas. Nuevos algoritmos también pueden ser desarrollados para aprovechar las distintas posibilidades de hardware digital, tal como apoyo para un rápido procesamiento en serie.

Las ventajas que la inteligencia artificial, hardware y software combinados, podrían alcanzar en última instancia, son enormes.[33] Pero ¿con qué rapidez podrían alcanzarse esas ventajas potenciales? Esa es la pregunta que abordaremos ahora.

CAPÍTULO 4

La cinética de una explosión de inteligencia

Una vez que las máquinas se equiparen con el ser humano en cuanto a su capacidad de razonamiento general, ¿cuánto tiempo pasará antes de que alcancen la superinteligencia radical? ¿Será ésta una transición lenta, gradual y prolongada? ¿O será repentina, explosiva? En este capítulo analizaremos la cinética de la transición a la superinteligencia como una función de la potencia de optimización y de la resistencia al progreso del sistema. Consideraremos lo que sabemos o podemos suponer razonablemente sobre el comportamiento de estos dos factores en el ámbito de la inteligencia general de nivel humano.

Sincronización y velocidad de despegue

Teniendo en cuenta que las máquinas superarán ampliamente *con el tiempo* a la biología en inteligencia general, pero que *actualmente* la

cognición artificial es mucho más restringida que la cognición humana, nos preguntamos cómo de rápido se llevará a cabo esta superación. La pregunta que aquí nos hacemos debe distinguirse claramente de la cuestión que consideramos en el capítulo 1 sobre lo lejos que actualmente estamos de desarrollar una máquina con inteligencia general de nivel humano. Aquí la pregunta es que, *contando con que dicha máquina se desarrolle, ¿cuánto tiempo pasará a partir de entonces hasta que una máquina se vuelva radicalmente superinteligen- te?* Téngase en cuenta que uno puede creer que las máquinas tardarán bastante tiempo en llegar a la línea de base humana, o uno puede ser agnóstico acerca de cuánto tiempo tardarán, y, sin embargo, tener una creencia firme en que, una vez que esto suceda, el posterior ascenso hasta la superinteligencia en sentido fuerte será muy rápido.

Puede ser útil pensar esquemáticamente sobre estos asuntos, a pesar de que hacerlo implique ignorar temporalmente algunas calificaciones y detalles complicados. Consideremos, pues, un diagrama que represente gráficamente la capacidad intelectual del sistema de inteligencia artificial más avanzado como una función en el tiempo (Figura 7).

Una línea horizontal denominada “línea de base humana” representa las capacidades intelectuales eficaces de un ser humano adulto representativo con acceso a las fuentes de información y ayudas tecnológicas disponibles en la actualidad en los países desarrollados. Actualmente, el sistema más avanzado de IA está muy por debajo

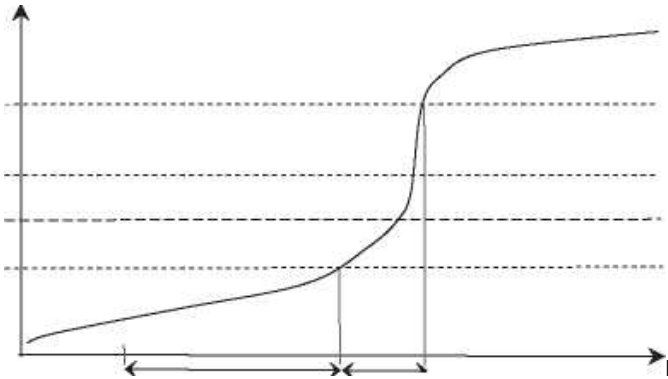


Figura 7. Configuración del despegue. Es importante distinguir entre las siguientes preguntas: "¿Se producirá un despegue? y si es así, ¿cuándo?" y "Si efectivamente se produce un despegue, ¿cómo de abrupto será?" Se podría sostener, por ejemplo, que va a pasar mucho tiempo antes de que ocurra un despegue, pero que cuando lo haga, sucederá rápidamente. Otra cuestión relevante (que no se ilustra en esta figura) es: "¿De qué magnitud será la fracción de la economía mundial que estará implicada en el despegue?" Estas preguntas están relacionadas pero son distintas.

de la línea de base humana tomando como referencia cualquier métrica razonable de la capacidad intelectual general. En algún momento en el futuro, una máquina podría llegar a una paridad aproximada respecto de esta línea de base humana (que postulamos estará fijada y anclada para el año 2014, por ejemplo, aunque las capacidades de los individuos humanos deberían haber aumentado en los años intermedios): esto marcaría el inicio del despegue. Las capacidades del sistema seguirían creciendo, y en algún momento posterior, el sistema alcanzaría la paridad respecto de la capacidad combinada intelectual de toda la humanidad (de nuevo anclado en el presente): lo que podríamos llamar la "línea de base de la civilización" Eventualmente, si las capacidades del sistema siguen creciendo, alcanzaría la "superinteligencia fuerte", un nivel de inteligencia muy superior al promedio intelectual combinado de la humanidad contemporánea. El logro de la superinteligencia fuerte marcaría la finalización del despegue, aunque el sistema pudiera seguir ganando en capacidad a partir de entonces. En algún momento durante la fase de despegue, el sistema podría pasar un punto de referencia que podemos llamar "el cruce", un punto más allá del cual una mejora adicional del sistema estaría impulsada principalmente por las propias acciones del sistema y no por el trabajo realizado en ella por otros.[1] (La posible existencia de un cruce como éste será importante en la

subsección sobre el poder de optimización y la explosividad, más adelante en este capítulo).

Con esta imagen en mente, podemos distinguir tres clases de escenarios de transición —escenarios en los que los sistemas evolucionan de un nivel de inteligencia humana hasta una superinteligencia— en función de su brusquedad; es decir, si representan un despegue lento, rápido, o moderado.

Lento

Un despegue lento es el que se produce en un intervalo temporal largo, como décadas o siglos. Los escenarios de despegue lento ofrecen excelentes oportunidades para que los procesos políticos humanos se adapten y respondan. Diferentes enfoques podrían ser juzgados y proba

dos poco a poco. Nuevos expertos podrían ser entrenados y acreditados. Campañas de base podrían ser movilizadas por grupos que sientan que están siendo perjudicados por el desarrollo de los acontecimientos. Si pareciera que nuevos tipos de infraestructuras más seguras o de vigilancia en masa para investigadores de IA fueran necesarias, se podrían desarrollar y desplegar tales sistemas. Naciones temerosas de una carrera armamentista de IA tendrían tiempo para tratar de negociar tratados y diseñar mecanismos que los aplicaran. La mayoría de las preparaciones realizadas antes del inicio del lento despegue irían quedando obsoletas a medida que mejores soluciones fueran haciéndose gradualmente visibles a la luz del amanecer de la nueva era.

Rápido

Un despegue rápido es aquel que se produce en un intervalo temporal corto, como minutos, horas o días. Los escenarios de despegue rápido ofrecen poca oportunidad para que los seres humanos puedan deliberar Nadie tiene por qué siquiera notar algo inusual antes de que la partida esté ya perdida. En un escenario de despegue rápido, el destino de la humanidad depende esencialmente de los preparativos previamente establecidos. En el extremo más lento dentro del rango de escenarios de despegue rápido, algunas acciones humanas simples podrían ser posibles, como abrir de un click la "maleta nuclear"; pero dicha acción tendría que ser o bien elemental, o bien tendría que haberse planificado y preprogramado con antelación.

Moderado

Un despegue moderado es aquel que se produce en un intervalo temporal intermedio, como meses o años. Los escenarios de despegue moderados dan a los humanos alguna oportunidad de responder, pero no mucho tiempo para analizar la situación, para probar diferentes enfoques, o para resolver complicados problemas de coordinación. No habría tiempo suficiente para desarrollar o implementar nuevos sistemas (por ejemplo, sistemas políticos, regímenes de vigilancia, o protocolos de seguridad en la red del ordenador), pero se podrían aplicar los sistemas existentes al nuevo desafío.

En un despegue lento, habría tiempo de sobra para que las noticias salieran a la luz. En un despegue moderado, por el contrario, es posible que los acontecimientos se mantuvieran en secreto mientras se desarrollaran. El conocimiento puede estar restringido a un pequeño grupo de iniciados, como en un programa de investigación militar patrocinado de manera encubierta por el Estado. Proyectos comerciales, pequeños equipos académicos, y grupos del tipo de “nueve hackers en un sótano” también podrían ser clandestinos; sin embargo, si la perspectiva de una explosión de inteligencia estuviera “en el radar” de los organismos de inteligencia del Estado como una prioridad de seguridad nacional, el más prometedor de los proyectos privados tendría muchas posibilidades de estar bajo vigilancia. El Estado donde se desarrollaran (o una potencia extranjera dominante) tendrían la opción de nacionalizar o cerrar cualquier proyecto que mostrara indicios de haber comenzado el despegue. Los despegues rápidos se suceden tan rápidamente que no habría mucho tiempo para que fueran conocidos o para que alguien llevara a cabo una reacción

significativa incluso si llegara a conocerlo. Pero un extraño al proyecto podría intervenir *antes* del despegue si creyera que un proyecto en particular estuviera acercándose al éxito.

Los escenarios de despegue moderados podrían conducir a una turbulencia geopolítica, social y económica, si los individuos y grupos preeminentes lograran posicionarse para beneficiarse de la transformación en curso. Tal agitación, en caso de producirse, podría obstaculizar los esfuerzos para orquestar una respuesta bien coordinada; alternativamente, podría permitir soluciones más radicales de lo que circunstancias más tranquilas permitirían. Por ejemplo, en un escenario de despegue moderado donde las emulaciones baratas y capaces u otras mentes digitales fueran inundando gradualmente los mercados de trabajo en unos años, uno podría imaginar protestas masivas por parte de los trabajadores despedidos que presionan a los gobiernos para aumentar las prestaciones por desempleo o que insisten en pedir una garantía de salario digno para todos los ciudadanos humanos, o en recaudar impuestos especiales, o en imponer salarios mínimos a los empresarios que utilicen a emulaciones como trabajadores. Para que cualquier alivio derivado de este tipo de políticas no fuera sólo pasajero, el apoyo tendría que estar cimentado de alguna manera en estructuras de poder permanente. Problemas similares pueden surgir si el despegue es lento y no moderado, pero el desequilibrio y el rápido cambio de escenarios moderados presentan oportunidades especiales para que grupos pequeños ejerzan una influencia desproporcionada.

Podría parecerle a algunos lectores que de estos tres tipos de escenario, el despegue lento es el más probable, el despegue moderado es el menos probable, y el despegue rápido es totalmente inverosímil. Podría parecer descabellado suponer que el mundo pudiera ser transformado radicalmente y que la humanidad pudiera ser derrocada de su posición como pensador alfa en el transcurso de una o dos horas. Ningún cambio de tal rapidez ha ocurrido nunca en la historia humana, y sus paralelos más cercanos —las Revoluciones Industrial y Agrícola— se desarrollaron a lo largo de escalas temporales mucho más amplias (en torno a siglos y milenios en el primer caso, en torno a décadas y siglos en el segundo). Así que la tasa base para el tipo de transición que entraña un escenario de despegue rápido o medio, en términos de velocidad y magnitud del cambio postulado, es cero: carece de precedente fuera del mito y la religión.[2]

Sin embargo, este capítulo presentará algunas razones para pensar que el escenario de transición lenta es improbable. Cuando se produzca el despegue, si se produce, es probable que sea explosivo.

Para comenzar a analizar la cuestión de cómo de rápido será el despegue, podemos concebir el ritmo de aumento de inteligencia en

un sistema como una función (monótonamente creciente) de dos variables: la cantidad de "poder de optimización", o esfuerzo de calidad del diseño, que se aplica para aumentar la inteligencia del sistema; y la capacidad de respuesta del sistema a la aplicación de una cantidad dada de tal poder de optimización. Podríamos llamar al inverso de la capacidad de respuesta "resistencia al progreso", y escribir:

$$\text{Ratio de cambios en inteligencia} = \frac{\text{Poder de optimización}}{\text{Resistencia al progreso}}$$

Sin ninguna especificación sobre cómo cuantificar la inteligencia, el esfuerzo de diseño, y la resistencia al progreso, esta expresión sería meramente cualitativa. Pero al menos podemos observar que la inteligencia de un sistema aumentará rápidamente si *o bien* se le aplica mucho esfuerzo específico para aumentar su inteligencia y la inteligencia del sistema no es demasiado contraria a ese aumento, *o bien* hay un esfuerzo de diseño no arbitrario y la resistencia al progreso del sistema es baja (o ambos). Si sabemos la cantidad de esfuerzo de diseño destinada a la mejora de un sistema en particular, y la tasa de mejora que este esfuerzo produce, podríamos calcular la resistencia al progreso del sistema.

Además, se puede observar que la cantidad de energía de optimización dedicada a mejorar el rendimiento de un sistema varía entre los sistemas y con el tiempo. La resistencia al progreso de un sistema también puede variar dependiendo de la cantidad del sistema que ya se haya optimizado. A menudo, las mejoras más fáciles se hacen primero, derivando en rendimientos decrecientes (aumentando la resistencia al progreso) como frutos maduros que menguan. Sin embargo, también puede haber mejoras que hagan más fácil otras mejoras adicionales, lo que conduce a cascadas de mejora. El proceso de resolver un rompecabezas comienza siendo sencillo —es fácil encontrar las esquinas y los bordes. Entonces la resistencia al progreso aumenta cuando vemos que las piezas posteriores son más difíciles de encajar. Pero a medida que el rompecabezas llega a su fin, el espacio de búsqueda va decayendo y el proceso se hace más fácil de nuevo.

Para continuar en nuestra investigación, debemos, por tanto, analizar cómo la resistencia al progreso y la optimización de potencia puede variar en los períodos críticos durante el despegue. Esto nos va a ocupar en las próximas páginas.

Resistencia al progreso

Comencemos por la resistencia al progreso. Las perspectivas aquí dependen del tipo del sistema en consideración. Por completitud, primero echemos un breve vistazo a la resistencia al progreso que se encuentra a lo largo de las rutas de acceso a la super- inteligencia que no implican la inteligencia artificial avanzada. Encontramos que la resistencia al progreso por esos caminos parece ser bastante alta. Una vez hecho esto, vamos a volver al caso principal, que el despegue implique inteligencia por parte de las máquinas; y allí nos encontramos con que la resistencia al progreso en el momento crítico parece baja.

Caminos que no implican la inteligencia artificial

La mejora cognitiva a través de mejoras en salud pública y dieta tienen un gran rendimiento decreciente.[3] Se obtienen muchos beneficios al eliminar las deficiencias nutri- cionales graves, y las deficiencias más graves ya se han eliminado en gran medida en todos los países excepto en los más pobres. Mediante el aumento de una dieta ya bastante buena sólo obtendríamos una figura más esbelta. La educación probablemente también sufriría rendimientos decrecientes. La fracción de individuos talentosos en el mundo que no tienen acceso a educación de calidad sigue siendo importante, pero está en declive.

Los potenciadores farmacológicos pueden ofrecer algunos beneficios cognitivos durante las próximas décadas. Pero después de que las correcciones más fáciles se hayan logrado —tal vez aumentos sostenibles en energía mental y en la capacidad de concentración, junto con un mejor control sobre la consolidación de la memoria a largo plazo— las ganancias subsiguientes serán cada vez más difíciles de conseguir. A diferencia de los enfoques dietéticos y de salud pública, sin embargo, la mejora de la cognición a través de drogas inteligentes podría empezar siendo más fácil para después volverse más problemática. El campo de la neurofarmacología todavía carece de gran parte de los conocimientos básicos necesarios para intervenir de forma competente en el cerebro sano. La desatención a la medicina mejorativa como área legítima para la investigación puede ser parte de la culpa de este retraso actual. Si la neurocien- cia y la farmacología continúan progresando durante más tiempo sin centrarse en la mejora cognitiva, entonces tal vez habría algunos aumentos relativamente sencillos que se conseguirían cuando por fin el desarrollo de nootrópicos se convirtiera en algo claramente prioritario.[4]

La mejora cognitiva genética tiene un perfil de resistencia al progreso en forma de U similar a la de los nootrópicos, pero con mayores ganancias potenciales. La resistencia al progreso empezaría en un nivel alto si el único método disponible fuera la cría selectiva sostenida a lo largo de muchas generaciones, algo que obviamente es difícil de lograr a escala mundial de manera significativa. La mejora genética será más fácil a medida que la tecnología desarrollada para pruebas genéticas y selección sea barata y eficaz (particularmente cuando la selección de embriones por iteración sea factible en seres humanos). Estas nuevas técnicas permitirán aprovechar la reserva de variaciones genéticas humanas existentes para mejorar los alelos relacionados con la inteligencia. Como los mejores alelos existentes quedarían incorporados en paquetes de mejoras genéticas, sin embargo, ulteriores ganancias serán más difíciles de conseguir. La necesidad de realizar enfoques más innovadores en lo referente a la modificación genética puede entonces aumentar la resistencia al progreso. Hay límites a lo rápido que se puede progresar en el camino de la mejora genética, sobre todo por el hecho de que las intervenciones en la línea germinal están sujetas a un retraso madurativo inevitable: esto contrarresta fuertemente la posibilidad de un despegue rápido o moderado.[5] El hecho de que la selección de embriones sólo se pueda aplicar en el contexto de la fertilización in vitro ralentizará su tasa de adopción: otro factor limitante.

La resistencia al progreso por el camino de la interfaz cerebro-ordenador parece inicialmente muy alta. En el improbable caso de que de alguna manera se convierta en algo fácil insertar implantes en el cerebro y lograr su integración funcional de alto nivel con la corteza, la resistencia al progreso podría caer en picado. A la larga, la dificultad para avanzar en este camino sería similar a la implicada en la mejora de emulaciones o de IAs, ya que la mayor parte de la inteligencia del sistema cerebro-ordenador residiría finalmente en la parte informática.

La resistencia al progreso para hacer redes y organizaciones más eficientes *en general* es alta. Una gran cantidad de esfuerzo estaría destinada a la superación de esta resistencia al progreso, y el resultado en la mejora de la capacidad total de la humanidad de tal vez no más de un dos por ciento anual.[6] Por otra parte, los cambios en el entorno interno y externo implican que las organizaciones, aunque fueran eficientes en un determinado momento, pronto estarían mal adaptadas a sus nuevas circunstancias. Por tanto, se requiere un esfuerzo continuo de reforma aunque sólo sea para evitar el deterioro. Un cambio de ritmo en la tasa de ganancia en eficiencia organizativa media es quizá concebible, pero es difícil ver cómo, incluso en el escenario más radical, podría producirse otra cosa que un despegue

lento, ya que las organizaciones gestionadas por seres humanos se limitan a trabajar en escalas de tiempo humanas. Internet sigue siendo una frontera emocionante con muchas oportunidades para la mejora de la inteligencia colectiva, con una resistencia al progreso que parece estar, por el momento, en un rango moderado —el progreso es más o menos rápido, pero requerirá mucho esfuerzo hacer que este progreso efectivamente tenga lugar. Se puede esperar que la resistencia al progreso aumente cuando las frutas maduras (como los motores de búsqueda y el correo electrónico) vayan agotándose.

Caminos a través de la emulación y la IA

La dificultad de avanzar hacia la emulación de cerebro completo es difícil de estimar. Sin embargo, podemos señalar un hito específico del futuro: la emulación exitosa del cerebro de un insecto. Ese hito se encuentra en una colina, y su conquista dejaría a la vista gran parte del terreno por delante, lo que nos permitiría hacer una conjetura decente de la resistencia al progreso que traería consigo aumentar esta tecnología hasta la emulación de cerebro completo humano. (Una emulación de éxito del cerebro de un mamífero pequeño, tal como el de un ratón, daría un mejor punto de observación y permitiría estimar la distancia que falta hasta una emulación total del cerebro humano con un alto grado de precisión). El camino hacia la inteligencia artificial, por el contrario, puede no presentar un hito o punto de observación temprano tan evidente. Es muy posible que la búsqueda de la inteligencia artificial parezca perderse en una densa selva hasta que un avance inesperado revele la línea de llegada en un claro a sólo unos pasos de distancia.

Recordemos la distinción entre estas dos preguntas: ¿Cómo de difícil es alcanzar niveles más o menos humanos de capacidad cognitiva? ¿Y cómo de difícil es llegar desde allí a niveles sobrehumanos? La primera pregunta es sobre todo relevante para predecir cuánto tiempo pasará antes del inicio del despegue. La segunda pregunta es la clave para evaluar la forma del despegue, que es nuestro objetivo ahora. Y aunque podría ser tentador suponer que el paso desde el nivel humano al nivel sobrehumano debe ser el más difícil —este paso, después de todo, tiene lugar “a mayor altitud”, donde la capacidad debe añadirse a un sistema ya muy capaz de por sí — esto sería una suposición muy arriesgada. Es muy posible que la resistencia al progreso *caiga* cuando una máquina alcance la paridad con el ser humano.

Consideremos en primer lugar la emulación de cerebro completo. Las dificultades para la creación de la primera emulación humana son de un tipo muy diferente de aquellas implicadas en la mejora de una

emulación existente. La creación de una primera emulación implica enormes desafíos tecnológicos, particularmente en lo que se refiere al desarrollo de las necesarias capacidades de escaneo y de interpretación de imágenes. Este paso también podría requerir cantidades considerables de capital físico —un parque mecánico de escala industrial con cientos de máquinas de escaneo de alto rendimiento no es inverosímil. Por contraste, la mejora de calidad de una emulación existente implica afinar algoritmos y estructuras de datos: es esencialmente un problema de software, y podría resultar ser mucho más fácil que el perfeccionamiento de la tecnología de imagen necesaria para crear la plantilla original. Los programadores pueden experimentar fácilmente con trucos como aumentar el número de neuronas en diferentes áreas corticales para ver cómo afecta al desempeño.[7] También podrían trabajar en la optimización de código y en la búsqueda de modelos computacionales más simples que preservaran la funcionalidad esencial de neuronas individuales o de pequeñas redes de neuronas. Si el último requisito tecnológico pendiente fuera el escaneado o la traducción, disponiendo de una potencia de cálculo relativamente abundante, entonces podría haber sucedido que no se hubiera prestado mucha atención durante la fase de desarrollo a la eficiencia de la puesta en práctica, y podrían estar disponibles oportunidades sencillas para el ahorro de eficiencia computacional. (Una reorganización arquitectónica más fundamental también podría ser posible, pero eso nos llevaría fuera del camino de la emulación y hacia el territorio de la IA).

Otra forma de mejorar el código base una vez que la primera emulación se hubiera producido es escanear cerebros adicionales con diferentes o superiores habilidades y talentos. También se produciría un crecimiento en la productividad como consecuencia de la adaptación a estructuras organizativas y flujos de trabajo de los atributos específicos de las mentes digitales. Puesto que no hay precedentes en la economía humana de un trabajador que pueda literalmente ser copiado, restablecido y puesto a funcionar a diferentes velocidades, los coordinadores de la primera población de emulaciones encontrarían mucho margen para innovar en prácticas de gestión.

Después de un primer momento de caída, cuando la emulación de cerebro humano completo se haga posible, la resistencia al progreso puede subir de nuevo. Tarde o temprano, las ineficiencias más evidentes de la puesta en práctica se habrán optimizado, las variaciones algorítmicas más prometedoras se habrán probado y las oportunidades más fáciles para la innovación organizacional se habrán aprovechado. La biblioteca de plantillas se habrá ampliado de forma que la adquisición de más escáneres cerebrales añadirá poco beneficio

al trabajo con las plantillas existentes. Ya que una plantilla se puede multiplicar, cada copia puede ser entrenada individualmente en un campo diferente, y esto se puede hacer a una velocidad electrónica, por lo que podría ser que el número de cerebros que tuvieran que ser escaneados con el fin de capturar la mayor parte del potencial de ganancia económico fuera pequeño. Posiblemente un solo cerebro sería suficiente.

Otra posible causa de la escalada de resistencia al progreso es la posibilidad de que las emulaciones o sus partidarios biológicos se organicen para apoyar normas que restrinjan el uso de trabajadores de emulación, limitando la copia de emulaciones, prohibiendo ciertos tipos de experimentación con mentes digitales, instituyendo derechos para los trabajadores y un salario mínimo para las emulaciones, etc. Es igualmente posible, sin embargo, que los acontecimientos políticos vayan en la dirección opuesta, lo que contribuiría a una caída en la resistencia al progreso. Esto puede suceder si una restricción inicial en el uso de mano de obra de emulación da paso a una explotación sin trabas, ya que la competencia aumentaría y los costes económicos y estratégicos del terreno moral se habrían despejado.

En cuanto a la inteligencia artificial (una máquina inteligente, no una emulación), la dificultad de elevar un sistema desde el nivel humano hasta la inteligencia sobrehumana mediante mejoras algorítmicas depende de los atributos del sistema en particular. Diferentes arquitecturas pueden tener una resistencia al progreso muy diferente.

En algunas situaciones, la resistencia al progreso podría ser extremadamente baja. Por ejemplo, si una IA de nivel humano se retrasara porque una idea clave se le escapara durante mucho tiempo a los programadores, entonces, cuando se produjera la ruptura definitiva, la IA podría saltar de forma radical por encima del nivel humano sin siquiera tocar los peldaños intermedios. Otra situación en la que la resistencia al progreso podría llegar a ser extremadamente baja es la de un sistema de inteligencia artificial que pudiera alcanzar la inteligencia a través de dos modos diferentes de procesamiento. Para ilustrar esta posibilidad, supongamos que una IA se compone de dos

subsistemas, uno que posee técnicas de resolución de problemas de ámbito específico, otro que posee capacidad de razonamiento general. Así, podría darse el caso de que mientras el segundo subsistema permaneciera por debajo de un cierto umbral de capacidad, éste no contribuiría en nada al rendimiento general del sistema, ya que las soluciones que genera son siempre inferiores a las generados por el subsistema de ámbito específico. Supongamos ahora que una pequeña cantidad de energía de optimización se aplicara al subsistema de propósito general y que esto produjera un aumento rápido en la capacidad de ese subsistema. En un primer momento, no se observaría ningún aumento en el rendimiento del sistema en su conjunto, lo que indicaría que la resistencia al progreso es alta. Entonces, una vez que la capacidad del subsistema de propósito general cruzara el umbral en el que sus soluciones comienzan a imponerse a los del subsistema de dominio específico, el rendimiento del sistema global de repente comenzaría a mejorar al mismo ritmo acelerado que el subsistema de uso general, incluso aunque la cantidad de energía de optimización aplicada permanezca constante: la resistencia al progreso del sistema caería en picado.

También es posible que nuestra tendencia natural a ver la inteligencia desde una perspectiva antropocéntrica nos lleve a subestimar las mejoras en los sistemas sub-hu- manos, y por lo tanto a sobreestimar la resistencia al progreso. Eliezer Yudkowsky, un teórico de la IA que ha escrito mucho sobre el futuro de la inteligencia artificial, presenta el asunto de la siguiente manera:

> Una IA podría realizar un salto en Inteligencia aparentemente agudo puramente como resultado del antropomorfismo, la tendencia humana a pensar en el "tonto del pueblo" y en "Einstein", como los extremos de la escala de inteligencia, en lugar de entenderlos como puntos casi indistinguibles en la escala de las mentes-en-general. Todo lo que sea más tonto que un ser humano tonto puede parecemos a nosotros simplemente como "tonto". Uno se imagina la "flecha de la IA" moviéndose de manera constante en la escala de la inteligencia, superando a los ratones y los chimpancés, con IAs que aún parecen "tontas" porque dichas IAs no pueden hablar con fluidez o escribir artículos científicos, y entonces, veríamos que la flecha de la IA atravesaría la pequeña brecha entre infra-idiota hasta supra-Einstein en el transcurso de un mes o algún período similarmente corto.[8] (Véase figura 8).

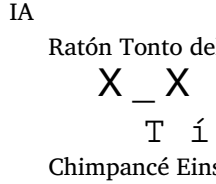


Figura 8. ¿Una escala menos antropomórfica? La diferencia entre un tonto y una persona inteligente puede parecer grande desde la perspectiva antropocéntrica, mas, en una visión menos provinciana, los dos tienen mentes casi indistinguibles.[9] Es casi seguro que resultará

más difícil y requerirá más tiempo construir una inteligencia artificial que tenga el nivel general de inteligencia del tonto del pueblo que mejorar un sistema de este tipo para que sea mucho más inteligente que cualquier humano.

El resultado de estas varias consideraciones es que es difícil predecir lo difícil que será hacer mejoras algorítmicas en la primera IA que alcance un nivel más o menos humano de inteligencia general. Hay por lo menos algunas circunstancias posibles en las que el algoritmo de resistencia al progreso es bajo. Pero incluso si el algoritmo de resistencia al progreso es muy alto, esto no excluye que la resistencia al progreso gene

ral de la IA sea baja. Para ello podría ser fácil de aumentar la inteligencia del sistema de otra manera que por la mejora de sus algoritmos. Hay otros dos factores que se pueden mejorar: el contenido y el hardware.

En primer lugar, tengamos en cuenta las mejoras de contenido. Por "contenido" aquí nos referimos a aquellas partes de los valores de software de un sistema que no componen su arquitectura algorítmica nuclear. Contenido puede incluir, por ejemplo, las bases de datos de percepciones almacenadas, las bibliotecas de habilidades especializadas, y los inventarios de conocimiento declarativo. Para muchos tipos de sistema, la distinción entre arquitectura algorítmica y contenido es muy borrosa; sin embargo, nos servirá como una manera simple y fácil de apuntar a una fuente potencialmente importante de ganancias de capacidad para una inteligencia artificial. Una forma alternativa de expresar la misma idea es diciendo que la capacidad de resolución de problemas intelectuales de un sistema se puede mejorar no sólo haciendo que el sistema sea más inteligente, sino también por la expansión de lo que el sistema sabe.

Consideremos un sistema de inteligencia artificial contemporáneo como el Text- Runner (un proyecto de investigación de la Universidad de Washington) o el Watson de IBM (el sistema que ganó el concurso *Jeopardy*!). Estos sistemas pueden extraer ciertas piezas de información semántica mediante el análisis de texto. Si bien estos sistemas no entienden lo que leen en el mismo sentido o en la misma medida en que lo hace un ser humano, no obstante pueden extraer cantidades significativas de información de lenguaje natural y utilizar esa información para hacer inferencias sencillas y contestar preguntas. También pueden aprender de la experiencia, construyendo representaciones más elaboradas de un concepto a medida que se encuentran con casos adicionales de su uso. Estos sistemas están diseñados para operar durante gran parte del tiempo sin supervisión (es decir, descifrando la estructura oculta de los datos no marcados en ausencia de señales de error o recompensa, sin la guía humana) y para ser rápidos y escalables. TextRunner, por ejemplo, trabaja con un corpus de 500 millones de páginas web.[10]

Ahora imaginemos un descendiente remoto de un sistema de este tipo que haya adquirido la capacidad de leer con la misma comprensión que un ser humano de diez años de edad, pero con una velocidad de lectura similar a la de TextRunner. (Esto es probablemente un problema de IA-completo). Así estaríamos imaginando un sistema que piensa mucho más rápido y tiene mucha mejor memoria que un adulto humano, pero que sabe mucho menos, y tal vez el efecto neto de esto sería que el sistema tendría más o menos la misma capacidad de resolución de problemas que un ser humano.

Pero su resistencia al progreso en cuanto a contenido es muy baja —lo suficientemente baja como para precipitar un despegue. En unas semanas, el sistema habría leído y dominado todo el contenido que figura en la Biblioteca del Congreso. Por entonces el sistema sabría mucho más que cualquier ser humano y pensaría muy rápido: se habría vuelto (por lo menos) levemente superinteligente.

Un sistema podría aumentar de este modo su capacidad intelectual efectiva en gran medida mediante la absorción de contenido pre-producido acumulado a través de siglos de ciencia y civilización humana: por ejemplo, mediante la lectura a través de internet. Si una IA alcanza el nivel humano sin haber tenido acceso a este material o sin haber sido capaz de digerirlo, entonces la resistencia al progreso general de la IA será baja aunque sea difícil mejorar su arquitectura algorítmica.

La resistencia al progreso por contenido también es un concepto relevante para las emulaciones. Una emulación de alta velocidad tiene ventaja no sólo porque pueda completar las mismas tareas que los humanos biológicos con mayor rapidez, sino porque también puede acumular contenido más oportuno, tal como habilidades y conocimientos relevantes para determinadas tareas. Con el fin de aprovechar todo el potencial de acumulación rápida de contenidos, sin embargo, un sistema debe tener en correspondencia una gran capacidad de memoria. No tiene mucho sentido leer una biblioteca entera si cuando se llega al águila ya se ha olvidado todo sobre la abeja. Mientras que un sistema de inteligencia artificial es probable que tenga una capacidad de memoria adecuada, las emulaciones heredarían algunas de las limitaciones de capacidad de sus plantillas humanas. Por lo tanto, podrían necesitar mejoras arquitectónicas para ser capaces de aprender sin límites.

Hasta ahora hemos considerado la resistencia al progreso de la arquitectura y de los contenidos, es decir, lo difícil que sería mejorar el software de inteligencia artificial que hubiera alcanzado la paridad con el ser humano. Ahora echemos un vistazo a una tercera forma de impulsar el rendimiento de la inteligencia artificial: la mejora de su hardware. ¿Cuál sería la resistencia al progreso de las mejoras enfocadas al hardware?

Comenzando con software inteligente (emulación o IA) puede amplificarse la *inteligencia colectiva* simplemente usando computadoras adicionales para ejecutar más instancias del programa.[11] También se podría amplificar la *inteligencia de velocidad* trasladando el programa a ordenadores más rápidos. Dependiendo del grado en que el programa se preste a la paralelización, la inteligencia de velocidad también podría ser amplificada mediante la ejecución del programa en más procesadores. Esto es probable que sea factible para emulaciones, que

tienen una arquitectura altamente paralelizada; pero muchos programas de IA tienen también importantes subrutinas que pueden beneficiarse de la paralelización masiva. Ampliar la *inteligencia de calidad* mediante el aumento de la potencia de cálculo también puede ser posible, aunque este caso sería menos directo.[12]

Por tanto, la resistencia al progreso para amplificar la inteligencia colectiva o la de velocidad (y, posiblemente, la inteligencia de calidad) en un sistema con software de nivel humano, probablemente sea baja. La única dificultad estaría en conseguir potencia de cálculo adicional. Hay varias maneras para que un sistema expanda su hardware de base, cada una relevante para diferentes escalas de tiempo.

En el corto plazo, la potencia de cálculo debería aumentar más o menos linealmente con la financiación: el doble de financiación compraría el doble de equipos, lo que permitiría el doble de instancias del software ejecutándose de forma simultánea. La aparición de servicios de computación en la nube ofrecería a un proyecto la opción de ampliar sus recursos computacionales sin tener que esperar a que los nuevos ordenadores sean entregados e instalados, aunque las preocupaciones sobre confidencialidad podrían favorecer el usar los ordenadores en casa. (En ciertas situaciones, la potencia de cálculo también podría ser obtenida por otros medios, como reclutando bots en internet[13]). Lo fácil que sea aumentar un determinado sistema por un factor depende de la cantidad de potencia de cálculo que utilice el sistema inicial. Un sistema que se ejecutara inicialmente en un PC podría aumentarse por un factor de miles con sólo un millón de dólares. Un programa ejecutado en un superordena- dor sería mucho más caro de aumentar.

En un plazo un poco más largo, el coste de adquirir hardware adicional puede elevarse cuando gran parte de las capacidades instaladas en el mundo se utilicen para ejecutar mentes digitales. Por ejemplo, en un escenario de emulación basada en el mercado competitivo, el coste de ejecutar una copia adicional de una emulación podría llegar a ser aproximadamente igual a los ingresos generados por la copia marginal, ya que los inversores habrían elevado el precio de las infraestructura informática existente para que coincidiera con el rédito que esperaban de su inversión (aunque si un solo proyecto alcanzara la tecnología, éste podría tener cierto grado de capacidad de monopolio en el mercado de la potencia de cálculo, y por lo tanto, pagar un precio más bajo).

En un plazo de tiempo un poco más grande, el suministro de potencia de cálculo crecerá a medida que las nuevas capacidades sean instaladas. Un aumento en la demanda estimularía la producción de fundiciones de semiconductores y estimularía la construcción de

nuevas plantas. (Un aumento particular de rendimiento, quizás de uno o dos órdenes de magnitud, también podría obtenerse mediante el uso de microprocesadores personalizados[14]). Por encima de todo, la creciente ola de mejoras tecnológicas proporcionará crecientes volúmenes de potencia de cálculo para las turbinas de las máquinas pensantes. Históricamente, la tasa de mejora en tecnología de la computación ha sido descrita por la famosa ley de Moore, que en una de sus variantes establece que la potencia de cálculo por dólar se duplica cada 18 meses más o menos.[15] Aunque no se puede estar seguro de que este ritmo de mejora continúe hasta el desarrollo de la inteligencia artificial de nivel humano, no dejará de haber lugar para avances en tecnología informática hasta que se alcancen límites físicos fundamentales.

Hay, pues, razones para esperar que la resistencia al progreso del hardware no sea muy alta. La consecución de más potencia de cálculo para un sistema, una vez que se compruebe su capacidad para conseguir una inteligencia de nivel humano, podría fácilmente suponer la adición de varios órdenes de magnitud en poder computacional (dependiendo de cómo de frugal respecto del hardware fuera el proyecto antes de la expansión). La configuración de chips puede añadir uno o dos órdenes de magnitud. Otros medios de ampliar la base del hardware, tales como la construcción de más fábricas o el progreso en las fronteras de la tecnología informática, tomarán más tiempo —normalmente varios años, a pesar de que este retraso se comprimirá radicalmente una vez que la superinteligencia artificial revolucione el desarrollo tecnológico y manufacturero.

En resumen, podemos hablar de la probabilidad de un *excedente de hardware*: cuando se cree software de nivel humano, es posible que ya esté disponible suficiente potencia computacional para ejecutar un gran número de copias a gran velocidad. La resistencia al progreso del software, como se mencionó anteriormente, es más difícil de calcular, pero podría ser incluso menor que la resistencia al progreso del hardware. En particular, podría existir un *excedente de contenido* en forma de contenidos previamente hechos (por ejemplo, internet) que estuvieran disponibles para ser utilizados por un sistema cuando alcanzara la paridad con lo humano. También es posible que haya *excedentes de algoritmos* —mejoras algorítmicas pre-diseñadas—, pero quizás es menos probable. Las mejoras de software (ya sean en forma de algoritmos o de contenido) podrían ofrecer mejoras de rendimiento de varios órdenes de magnitud que podrían ser fácilmente accesibles una vez que las mentes digitales alcancen la paridad con el ser humano, además de las mejoras de rendimiento alcanzables mediante el uso de más o mejor hardware.

Potencia de optimización y explosividad

Habiendo examinado la cuestión de la resistencia al progreso ahora debemos recurrir a la otra mitad de la ecuación esquemática, la *potencia de optimización.* Para refrescar la memoria: *Tasa de cambio de Inteligencia = potencia de optimización / resistencia al progreso.* Como se refleja en este esquema, un despegue rápido no requiere que la resistencia al progreso en la fase de transición sea baja. Un despegue rápido también podría suceder si la resistencia al progreso es constante o incluso moderadamente creciente, siempre que la potencia de optimización enfocada a mejorar el rendimiento del sistema crezca con suficiente rapidez. Como veremos a continuación, hay buenas razones para pensar que la potencia optimización aplicada *aumentará* durante la transición, al menos en ausencia de medidas deliberadas para evitar que esto suceda.

Podemos distinguir dos fases. La primera fase comienza con el inicio del despegue, cuando el sistema alcanza la línea base de la inteligencia humana individual. Como la capacidad del sistema sigue en aumento, podría usar parte o la totalidad de esa capacidad para mejorarse a sí misma (o para diseñar un sistema sucesor que, a nuestros efectos, viene a ser lo mismo). Sin embargo, la mayor parte de la potencia de optimización aplicada al sistema aún provendría de fuera del sistema, ya sea del trabajo de los programadores e ingenieros que trabajan en el proyecto o del tipo de trabajo realizado por el resto del mundo que pueda ser incorporado y utilizado por el proyecto.[16] Si esta fase se prolonga por un período significativo de tiempo, podemos esperar que la cantidad de energía aplicada en la optimización del sistema crezca. Las aportaciones tanto desde dentro del proyecto como desde el mundo exterior son propensas a aumentar a medida que las probabilidades de éxito del enfoque elegido se hacen manifiestas. Los investigadores trabajarían más duro, se reclutarían más investigadores y se compraría más potencia de cálculo para acelerar el progreso. El aumento podría ser especialmente dramático si el desarrollo de la inteligencia artificial de nivel humano tomara al mundo por sorpresa, en cuyo caso lo que antes hubiera sido un pequeño proyecto de investigación, pronto podría convertirse en el foco de intensos esfuerzos de investigación y desarrollo por parte de todo el mundo (aunque algunos de esos esfuerzos podrían ser canalizados hacia proyectos de la competencia).

Una segunda fase de crecimiento se iniciará si en algún momento el sistema adquiere tanta capacidad como para que la mayor parte de la potencia de optimización ejercida en él venga del propio sistema (marcado por el nivel variable denominado “cruce” en la Figura 7). Esto cambiaría fundamentalmente la dinámica, ya que cualquier aumento en la capacidad del sistema ahora se traduciría en un aumento proporcional de la cantidad de energía que se aplica a la

optimización de sus ulteriores mejoras. Si la resistencia al progreso permaneciera constante, esta retroalimentación dinámica produciría un crecimiento exponencial (véase el cuadro 4). Esta constante duplicadora depende del escenario, pero podría ser muy corta —de unos pocos segundos en algunos escenarios— si el crecimiento se produjera a velocidades electrónicas, que pueden acontecer como resultado de mejoras algorítmicas o de la explotación de un excedente de contenidos o de hardware.[17] El crecimiento impulsado por construcción física, tal como la producción de nuevos ordenadores o equipo de fabricación, requeriría una escala de tiempo algo más grande (pero todavía una que podría ser muy corta en comparación con el actual ritmo de crecimiento de la economía mundial).

del punto de cruce contribuye con fuerza a hacer que el despegue sea más rápido de lo que habría sido de otro modo.

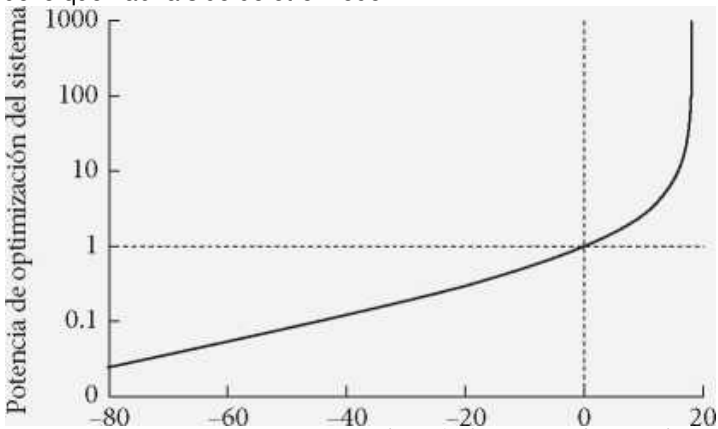


Figura 9. Un sencillo modelo de una explosión de inteligencia.

Por lo tanto, es probable que la potencia de optimización aplicada aumentara durante la transición: inicialmente, porque los humanos se esforzarían más en mejorar la inteligencia de la máquina que estuviera mostrando un potencial espectacular, más tarde, porque la propia inteligencia artificial se volvería capaz de impulsar un mayor progreso a velocidades digitales. Esto crearía una posibilidad real de despegue rápido o medio, *incluso si la resistencia al progreso fuera constante o ligeramente creciente en toda la línea base humana.*[18] No obstante, vimos en el apartado anterior que hay factores que podrían conducir a una gran caída en la resistencia al progreso en torno a la capacidad de línea base humana. Estos factores incluyen, por ejemplo, la posibilidad de expansión rápida de hardware una vez que se haya alcanzado una mente-software de trabajo; la posibilidad de mejoras algorítmicas; la posibilidad de escanear cerebros adicionales (en el caso de la emulación de cerebro completo); y la posibilidad de incorporar rápidamente grandes cantidades de contenido al digerir internet (en el caso de la inteligencia artificial).[24]

A pesar de estas observaciones, la forma de la curva de resistencia al progreso en la región clave aún no está bien caracterizada. En particular, no está claro lo difícil que sería mejorar la calidad del software de emulación de nivel humano o de la IA. Tampoco está clara la dificultad que entrañaría la ampliación de potencia de hardware disponible para un sistema. Mientras que hoy en día sería relativamente fácil aumentar la potencia computacional de un proyecto pequeño mediante el gasto de miles de veces más dinero en potencia de cálculo, o esperando unos años para que el precio de las

computadoras cayera, es posible que la primera inteligencia artificial que llegue a la línea de base humana sea el resultado de un gran proyecto que involucre a super- computadoras costosas, que no se puedan aumentar de escala a bajo precio, y que la ley de Moore haya expirado para entonces. Por estas razones, aunque un despegue rápido o medio parece más probable, la posibilidad de un despegue lento no puede ser desechada.[25]

CAPÍTULO 5

Ventaja estratégica decisiva

Una pregunta distinta pero relacionada con la cuestión de la cinética es ¿habrá un poder superinteligente o muchos? ¿Podría una explosión de inteligencia impulsar un proyecto tan adelantado respecto del resto como para que fuera capaz de dictar el futuro? ¿O será el progreso más uniforme, desplegándose en un frente amplio, con muchos proyectos participando pero sin que ninguno se asegure una ventaja abrumadora y permanente?

En el capítulo anterior se analizó un parámetro clave para determinar el tamaño de la brecha que podría abrirse entre un poder puntero y sus competidores más cercanos —a saber, la velocidad de transición desde una inteligencia humana a una inteligencia sobrehumana en sentido fuerte. Esto sugiere un primer análisis a bocajarro. Si el despegue fuera *rápido* (completado en el transcurso de horas, días o semanas), entonces es poco probable que los dos proyectos independientes estuvieran despegando al mismo tiempo: casi con certeza, el primer proyecto habría completado su despegue antes de que ningún otro proyecto hubiera comenzado su propio despegue. Si el despegue fuera *lento* (prolongándose durante muchos años o décadas), entonces podría haber múltiples proyectos despegando simultáneamente, de modo que aunque los proyectos hacia el final de la transición hubieran ganado enormemente en capacidad, no habría ningún momento en el que algún proyecto llegara a estar lo suficientemente lejos delante de los demás como para tener una ventaja abrumadora. Un despegue de velocidad *moderada* se posicionaría entre los anteriores, con una de estas posibilidades como condición: que pudiera haber o no más de un proyecto despegando al mismo tiempo.[1]

¿Llegaría un proyecto de inteligencia artificial a adelantarse tanto a la competencia como para conseguir una *ventaja estratégica decisiva* —es decir, un nivel de ventajas tecnológicas y otro tipo de ventajas suficientes como para alcanzar la dominación completa del mundo? ¿Si un proyecto obtuviera una ventaja estratégica decisiva, sería utilizada para eliminar a los competidores y formar una *Unidad* (un orden mundial en el que sólo existiría una agencia de toma de decisiones a nivel mundial)? Y si hubiera un proyecto ganador, ¿Cómo de “grande” sería —no en términos de tamaño físico o presupuesto, sino en términos de cuántos deseos personales estarían incorporados en su diseño? Abordaremos estas cuestiones de una en una.

¿Conseguirá el proyecto adelantado una ventaja estratégica decisiva?

Un factor que influye en el tamaño de la brecha entre el proyecto adelantado y sus perseguidores es la velocidad de difusión de aquello que da al líder una ventaja competitiva. Un proyecto adelantado podría tener dificultades para obtener y mantener una ventaja grande si los perseguidores pudieran copiar fácilmente sus ideas e innovaciones. La imitación crea un viento de frente que perjudica a los líderes y beneficia a los rezagados, especialmente si la propiedad intelectual está débilmente protegida. Un proyecto adelantado también puede ser especialmente vulnerable a la expropiación, a la tributación, o a la descomposición bajo regulaciones antimonopolio.

Sería un error, sin embargo, suponer que este viento de frente aumentará monótonamente a medida que crezca la brecha entre el proyecto adelantado y sus perseguidores. Así como un ciclista que se descuelga a demasiada distancia del pelotón ya no está protegido del viento por los ciclistas de delante, de igual modo un perseguidor tecnológico que se quedara bastante rezagado respecto de la primera línea podría encontrar dificultades para asimilar los avances que se realizaran en los primeros puestos.[2] La brecha en comprensión y capacidad podría haber crecido demasiado. El líder podría haber migrado a una plataforma tecnológica más avanzada, por lo que las innovaciones posteriores serían intransferibles a las plataformas primitivas utilizadas por los rezagados. Un líder lo suficientemente destacado podría tener la capacidad de frenar la fuga de información de sus programas de investigación y de sus instalaciones más delicadas, o la capacidad de sabotear los esfuerzos de sus competidores por desarrollar sus propias capacidades avanzadas.

Si el proyecto adelantado fuera un sistema de inteligencia artificial, podría tener atributos que harían más fácil la expansión de sus capacidades al tiempo que reducirían la velocidad de difusión. En las organizaciones humanas, las economías de escala están contrarrestadas por las ineficiencias burocráticas y problemas administrativos, incluyendo dificultades como la de mantener el secreto comercial.[3] Estos problemas presumiblemente limitarían el crecimiento de un proyecto de inteligencia artificial siempre que fuera manejado por seres humanos. Un sistema de inteligencia artificial, sin embargo, podría evitar algunas de estas ineficiencias de escala, ya que los módulos de la IA (en contraste con los de los trabajadores humanos) no tienen por qué tener preferencias individuales que difieran de los del sistema en su conjunto. Así, el sistema de inteligencia artificial podría evitar una parte considerable de las ineficiencias derivadas de los problemas administrativos de las

empresas humanas. La misma ventaja —tener partes completamente leales— también haría que fuera más fácil para un sistema de inteligencia artificial conseguir metas clandestinas de largo alcance. Una IA no tendría empleados descontentos predispuestos a ser captados por los competidores o a ser sobornados para convertirse en informadores.[4]

Podemos tener una idea de la distribución de las posibles brechas temporales de desarrollo observando algunos ejemplos históricos (véase el cuadro 5). Parece que lo habitual para proyectos tecnológicos estratégicamente importantes está entre unos pocos meses y unos pocos años.

Es posible que la globalización y el aumento de la vigilancia reduzcan los típicos retrasos entre proyectos tecnológicos en competencia. Sin embargo, no es probable que se acorte el promedio de los retrasos (en ausencia de coordinación deliberada).[21]

disponible; excepto en algunos casos, cuando parece que el descubrimiento ha de ofrecer una ventaja estratégica, y la publicación se retrasa. Por ejemplo, dos de las ideas más importantes de la criptografía de clave pública son el protocolo de intercambio de claves Diffie-Hellman y el esquema de cifrado RSA. Estos fueron descubiertos por la comunidad académica en 1976 y 1978 respectivamente, pero posteriormente se ha confirmado que los criptógrafos del grupo de seguridad de comunicaciones del Reino Unido las conocían desde principios de los años setenta.[20] Los grandes proyectos de software podrían ofrecer una analogía más estrecha con los proyectos de IA, pero es más difícil dar ejemplos nítidos de retrasos típicos ya que el software se suelen dar a conocer de manera incremental y las funcionalidades de los sistemas de la competencia a menudo no son directamente comparables.

Incluso en ausencia de dinámicas que condujeran a un efecto de bola de nieve, algunos proyectos acabarían naturalmente con un mejor personal de investigación, con mayor liderazgo, y con mejor infraestructura, o simplemente se tropezarían con mejores ideas. Si dos proyectos persiguen enfoques alternativos, uno de los cuales resulta funcionar mejor, es posible que los proyectos rivales tarden muchos meses en cambiar al enfoque superior incluso aunque fueran capaces de seguir de cerca lo que el líder estuviera haciendo.

Combinando estas observaciones con nuestra discusión anterior sobre la velocidad de despegue, se puede concluir que es muy poco probable que dos proyectos lleguen a estar lo suficientemente cerca como para hacer un despegue rápido al mismo tiempo; en un despegue medio, fácilmente podría darse de cualquier manera; y en un despegue lento, es muy probable que varios proyectos llevaran a cabo el proceso en paralelo. Pero el análisis necesita ir un paso más allá. La pregunta clave no es el número de proyectos que despegarían en tándem, sino cuántos proyectos emergerían lo suficientemente bien capacitados para que ninguno de ellos tuviera una ventaja estratégica decisiva. Si el proceso de despegue fuera relativamente lento al principio y luego se acelerara, la distancia entre los proyectos que compiten tendería a crecer. Regresando a nuestra metáfora de la bicicleta, la situación sería análoga a la de un par de ciclistas pedaleando en una colina empinada, uno detrás del otro a cierta distancia —la brecha entre ellos crecería cuando el ciclista adelantado superara la cima y comenzara a acelerar por la bajada del otro lado.

Considérese el siguiente escenario de despegue medio. Supongamos que se necesitara un proyecto de un año para aumentar la capacidad de una IA desde la línea base humana hasta una superinteligencia fuerte, y que el proyecto entrara en fase de despegue con una ventaja de seis meses respecto del próximo proyecto más avanzado. Los dos proyectos despegarían simultáneamente. Podría parecer, entonces, que ningún proyecto llegaría a tener una ventaja

estratégica decisiva. Pero no tiene por qué ser así. Supongamos que se necesitaran nueve meses para avanzar desde la línea base humana hasta el punto de cruce, y otros tres meses a partir de ahí hasta la superinteli- gencia fuerte. El proyecto adelantado entonces alcanzaría una superinteligencia fuerte tres meses antes de que el siguiente proyecto siquiera llegara al punto de cruce. Esto

daría al proyecto líder una ventaja estratégica decisiva y la oportunidad de valerse de su liderazgo para obtener el control permanente mediante la desactivación de los proyectos en competencia y el establecimiento de una Unidad. (Nótese que el concepto de Unidad es uno abstracto: una Unidad podría ser una democracia, una tiranía, una sola IA dominante, un sólido conjunto de normas globales que incluyeran disposiciones eficaces para su cumplimiento, o incluso un gobernante supremo alienígena —su definición característica es simplemente que es algún tipo de administración que puede resolver los principales problemas de coordinación global. Podría, aunque no es necesario, recordar a alguna forma consabida de gobierno humano[22]).

Dado que no hay una perspectiva especialmente clara de crecimiento explosivo justo después del punto de cruce, cuando el fuerte circuito de retroalimentación positiva de la potencia optimizadora entre en juego, un escenario de este tipo es una posibilidad seria, y aumenta las posibilidades de que el proyecto que lleve la delantera alcance una ventaja estratégica decisiva, aunque el despegue no fuera rápido.

¿Qué magnitud tendrá el proyecto exitoso?

Algunas rutas de acceso a la superinteligencia requieren de grandes recursos y, por lo tanto, es probable que sean exclusivamente proyectos grandes y bien financiados. La emulación de cerebro completo, por ejemplo, requiere diferentes tipos de conocimientos y un montón de equipo. Las mejoras en inteligencia biológica y en interfaces cerebro-ordenador también tendrían un factor de gran escala: mientras que una pequeña empresa de biotecnología podría inventar una o dos drogas, logrando la super- inteligencia a través de una de estas rutas (si fuera factible en absoluto) probablemente requerirá muchos inventos y muchas pruebas, y, por lo tanto, necesitará el respaldo de un sector industrial o de un programa nacional bien financiado. Lograr la super- inteligencia colectiva haciendo más eficientes las organizaciones y redes requiere una financiación aún más extensa, que incluiría la participación de gran parte de la economía mundial.

El camino de la IA es más difícil de evaluar. Tal vez sería necesario un programa de investigación muy grande; tal vez podría ser realizado por un grupo pequeño. Un escenario de pirata informático solitario tampoco puede excluirse. La construcción de una IA seminal podría requerir conocimientos y algoritmos desarrollados a lo largo de muchas décadas por la comunidad científica mundial. Pero es posible

que la última idea crítica en el avance pudiera provenir de una sola persona o de un pequeño grupo que consiguiera encajarlo todo. Este escenario es menos realista para algunas arquitecturas IA que para otras. Un sistema que tuviera un gran número de piezas que necesitaran ser ajustadas y afinadas para funcionar juntas de manera efectiva, y que luego necesitara ser cuidadosamente cargado de contenido cognitivo a medida, es probable que requiriera un proyecto más amplio. Pero si una IA seminal pudiera ser el modelo de un sistema simple, uno cuya construcción dependiera sólo de alcanzar unos principios básicos adecuados, entonces la hazaña podría estar al alcance de un pequeño equipo o de un individuo. La probabilidad de que el avance final sea llevado a cabo por un pequeño proyecto aumenta si la mayor parte del progreso anterior en dicho campo se halla publicado de manera abierta o está disponible como software de código abierto.

Hay que distinguir la pregunta sobre lo grande que será el proyecto que directamente *construya* el sistema, de la pregunta sobre lo grande que será el grupo que *controle* si, cómo y cuándo se crea el sistema. La bomba atómica fue creada principalmente por un grupo de científicos e ingenieros. (El Proyecto Manhattan empleó a alrededor de 130.000 personas en su apogeo, la gran mayoría de los cuales eran trabajadores de la construcción u obreros[23]). Estos expertos técnicos, sin embargo, fueron controlados por el ejército estadounidense, que a su vez estaba dirigido por el gobierno de los Estados Unidos, que en última instancia era responsable ante el electorado estadounidense, que en su momento constituía aproximadamente una décima parte de la población adulta mundial.[24]

Monitorización

Dadas las implicaciones de seguridad extrema de la superinteligencia, los gobiernos probablemente traten de nacionalizar cualquier proyecto de su territorio que ellos consideren cercano a lograr un despegue. Un Estado poderoso también podría intentar adquirir proyectos ubicados en otros países a través del espionaje, el robo, el secuestro, el soborno, las amenazas, la conquista militar, o cualquier otro medio disponible. Un Estado de gran alcance que no pudiera adquirir un proyecto extranjero podría, en su lugar, destruirlo, sobre todo si el país receptor careciera de un medio eficaz de disuasión. Si las estructuras de gobierno mundial fueran ya fuertes en el momento en que un avance comenzara a parecer inminente, es posible que los proyectos prometedores fueran puestos bajo control internacional.

Una cuestión importante, por lo tanto, es si las autoridades

nacionales o internacionales serían capaces de ver venir una explosión de inteligencia inminente. En la actualidad, las agencias de inteligencia no parecen estar buscando con mucho ahínco proyectos de IA prometedores u otras formas de amplificación de la inteligencia.[25] Si de hecho no prestan (mucha) atención, esto se debe probablemente a la percepción ampliamente compartida de que no hay ninguna posibilidad de que surja una superinteligencia de manera inminente. Si llegara a convertirse en una creencia común entre los científicos prestigiosos el que hay una posibilidad sustancial de que la superinteligencia estuviera a la vuelta de la esquina, las principales agencias de inteligencia del mundo probablemente comenzarían a monitorizar grupos e individuos que parecieran estar dedicados a investigaciones relevantes. Cualquier proyecto que comenzara a mostrar un progreso suficiente podría entonces ser nacionalizado rápidamente. Si las élites políticas se persuadieran de la gravedad del riesgo, los esfuerzos civiles en áreas delicadas podrían ser regulados o prohibidos.

¿Cómo de difícil será esa monitorización? La tarea es más fácil si el objetivo es sólo no perder de vista al proyecto puntero. En ese caso, una vigilancia centrada en los proyectos mejor dotados de recursos puede ser suficiente. Si el objetivo es, en cambio, evitar que cualquier trabajo tenga lugar (al menos fuera de las instituciones especialmente autorizadas), entonces la vigilancia tendría que ser más amplia, ya que muchos proyectos pequeños e individuos estarán en condiciones de hacer al menos algunos progresos.

Sería más fácil monitorizar proyectos que requieren de grandes cantidades de capital físico, como sería el caso de un proyecto de emulación de cerebro completo. La investigación de la inteligencia artificial, por el contrario, sólo requeriría un ordenador personal y, por lo tanto, sería más difícil de controlar. Parte del trabajo teórico se podría hacer con papel y lápiz. Aun así, no sería demasiado difícil de identificar a las personas más capaces, con un serio y documentado interés en investigaciones sobre inteligencia artificial general. Estas personas suelen dejar rastros visibles. Es posible que hayan publicado artículos académicos, presentado en congresos, publicado en foros de internet, o conseguido puestos en los principales departamentos de ciencia informática. También pueden haber tenido comunicaciones con otros investigadores de IA, lo que nos permitiría identificarlos mediante mapeados de gráfica social.

Los proyectos diseñados desde un principio para ser secretos podrían ser más difíciles de detectar. Un proyecto común de desarrollo de software podría servir como avanzadilla.[26] Sólo un cuidadoso análisis del código que estuvieran produciendo revelaría la verdadera naturaleza de lo que el proyecto estaría tratando de lograr. Tal análisis

requeriría una gran cantidad de mano de obra (altamente cualificada), por lo que sólo un pequeño número de proyectos sospechosos de este tipo podrían ser analizados. La tarea sería mucho más fácil si una tecnología de detección de mentiras efectiva se hubiera desarrollado y pudiera ser utilizada rutinariamente en este tipo de vigilancia.[27]

Otra razón por la que los Estados podrían dejar escapar acontecimientos precursores es la dificultad inherente de la previsión de algunos tipos de avance. Esto es más relevante para la investigación en IA que para el desarrollo de la emulación de cerebro completo, ya que este último es más probable que se vea precedido por una clara concatenación de avances clave.

También es posible que las agencias de inteligencia y otras administraciones gubernamentales sean algo torpes o rígidas y que esto les impida comprender el significado de algunos acontecimientos que podrían ser evidentes para algunos grupos externos. Las dificultades para la comprensión oficial de una potencial explosión de inteligencia podrían ser especialmente graves. Es concebible, por ejemplo, que el tema se inflame con controversias religiosas o políticas, convirtiéndolo en un tabú para los funcionarios de algunos países. El tema podría llegar a ser asociado con alguna figura desacreditada o con la charlatanería y la publicidad en general, y, por lo tanto, rechazado por los científicos respetados y otras figuras de la clase dirigente. (Como vimos en el capítulo 1, algo similar ya ha sucedido dos veces: recordemos los dos “inviernos de IA”). Los grupos industriales podrían ejercer presión política para evitar calumnias sobre áreas de negocio rentables; las comunidades académicas podrían cerrar filas para marginar a los que expresaran sus preocupaciones acerca de las consecuencias a largo plazo de la ciencia que se estuviera haciendo.[28]

En consecuencia, un fracaso total de la inteligencia no puede descartarse. Tal falta es especialmente probable si los avances se produjeran en el futuro más próximo, antes de que el tema se haya elevado al debate público. E incluso si las agencias de inteligencia no se equivocan, los líderes políticos podrían no escuchar o actuar según sus consejos. Dar comienzo al Proyecto Manhattan requirió un esfuerzo extraordinario por parte de varios físicos visionarios, incluyendo especialmente a Mark Oli- phant y Leó Szilárd: este último convenció a Eugene Wigner de que convenciera a Albert Einstein para poner su nombre en una carta dirigida a convencer al presidente Franklin D. Roosevelt de que investigara en el tema. Incluso después de que el proyecto llegara a su máximo nivel, Roosevelt se mantuvo escéptico respecto de su viabilidad e importancia, al igual que su sucesor Harry Truman.

Para bien o para mal, probablemente sería más difícil que un

pequeño grupo de activistas influyeran en una explosión de inteligencia si grandes jugadores, como los Estados, estuvieran tomando parte activa. Las oportunidades para que individuos particulares reduzcan la cantidad global de riesgo existencial de una potencial explosión de inteligencia son, por tanto, mayores en escenarios en los que los grandes jugadores siguen siendo relativamente ajenos a la cuestión, o en los que los primeros esfuerzos de los activistas marquen una gran diferencia sobre la posibilidad, el cuándo, los medios, o la actitud con que los grandes jugadores entren en el juego. Por lo tanto, los activistas que busquen el máximo impacto tal vez deseen centrar la mayor parte de su planificación en tales escenarios de gran influencia, incluso si creen que los escenarios en los que los grandes jugadores terminan acertando son más probables.

Colaboración Internacional

La coordinación internacional es más probable si las estructuras de gobierno global se hacen más fuertes. La coordinación también podría ser más probable si el significado de una explosión de inteligencia es muy apreciado antes de tiempo y si la supervisión efectiva de todos los proyectos serios es factible. Incluso si la monitorización no es factible, no obstante, la cooperación internacional seguiría siendo posible. Muchos países podrían unirse para apoyar un proyecto conjunto. Si tal proyecto conjunto tuviera recursos suficientemente buenos, podría tener una buena oportunidad de ser el primero en llegar a la meta, sobre todo si cualquier proyecto rival se ve obligado a ser pequeño y secreto para eludir la detección.

Existen precedentes de exitosas colaboraciones científicas multinacionales a gran escala, como la Estación Espacial Internacional, el Proyecto Genoma Humano y el Gran Colisionador de Hadrones.[29] Sin embargo, la principal motivación para la colaboración en esos casos era el de los costes compartidos. (En el caso de la Estación Espacial Internacional, el fomento de un espíritu de colaboración entre Rusia y los Estados Unidos fue un objetivo importante en sí mismo[30]). El logro de una colaboración similar en un proyecto con enormes implicaciones en seguridad sería más difícil. Un país que se creyera capaz de lograr un avance unilateral podría tener la tentación de ir por su cuenta en vez de subordinar sus esfuerzos a un proyecto conjunto. Un país también podría abstenerse de unirse a una colaboración internacional por temor a que otros participantes pudieran desviar ideas generadas en colaboración y utilizarlas para acelerar un proyecto nacional encubierto.

Un proyecto internacional necesitaría, por lo tanto, sobreponerse a

los principales problemas de seguridad, y probablemente se necesitaría una buena cantidad de confianza para ponerlo en marcha, una confianza que puede tardar tiempo en desarrollarse. Consideremos que incluso después del deshielo en las relaciones entre Estados Unidos y la Unión Soviética tras el ascenso de Gorbachov al poder, los esfuerzos en reducción de armas —algo en lo que ambas superpotencias estaban muy interesadas— tuvieron un comienzo irregular. Gorbachov deseaba reducciones drásticas en el armamento nuclear, pero las negociaciones se estancaron por la Iniciativa de Defensa Estratégica de Reagan ("Star Wars"), a la que el Kremlin se opuso enérgicamente. En la Cumbre de Reykjavik de 1986, Reagan propuso que Estados Unidos compartiera con la Unión Soviética la tecnología que se desarrollaba bajo la Iniciativa de Defensa Estratégica, por lo que ambos países podrían protegerse de los lanzamientos accidentales y defenderse frente a las naciones más pequeñas que pudieran desarrollar armas nucleares. Sin embargo, Gorbachov no fue persuadido por esta aparente propuesta inmejorable. Consideraba la táctica como un ardid, negándose a creer que los estadounidenses fueran a compartir los frutos de su investigación militar más avanzada en un momento en que ni siquiera estaban dispuestos a compartir con los soviéticos su tecnología para ordeñar vacas.[31] Independientemente de si Reagan fue, de hecho, sincero en su oferta de colaboración entre superpotencias, la desconfianza hizo que la propuesta no pudiera ver la luz.

La colaboración es más fácil de lograr entre aliados, pero incluso allí no es automática. Cuando la Unión Soviética y los Estados Unidos se aliaron contra Alemania durante la Segunda Guerra Mundial, Estados Unidos ocultó su proyecto de bomba atómica a la Unión Soviética. Los Estados Unidos sí colaboraron en el Proyecto Manhattan con Gran Bretaña y Canadá.[32] Del mismo modo, el Reino Unido ocultó a la Unión Soviética su éxito al descifrar el código alemán Enigma, pero lo compartió —aunque no sin reticencias— con Estados Unidos.[33] Esto sugiere que a fin de lograr colaboración internacional sobre alguna tecnología de importancia fundamental para la seguridad nacional, podría ser necesario haber construido previamente una relación cercana y de confianza.

Volveremos en el capítulo 14 sobre la conveniencia y viabilidad de la colaboración internacional en el desarrollo de tecnologías de amplificación de la inteligencia.

De una ventaja estratégica decisiva a la Unidad

¿Decidiría un proyecto que obtuviera una ventaja estratégica decisiva

usarla para formar una Unidad?

Consideremos una situación histórica vagamente análoga. Los Estados Unidos desarrollaron armas nucleares en 1945. Fue la única potencia nuclear en solitario hasta que la Unión Soviética desarrolló la bomba atómica en 1949. Durante este intervalo —y por algún tiempo más— los Estados Unidos pudieron haber tenido, o estuvieron en condiciones de lograr, una ventaja militar decisiva.

Los Estados Unidos, entonces, teóricamente podrían haber utilizado su monopolio nuclear para crear una Unidad. Una forma en la que podrían haberlo hecho habría sido embarcándose en un enorme esfuerzo por construir un gran arsenal nuclear y amenazando (y si es necesario, llevando a cabo) un primer ataque nuclear para destruir la capacidad industrial de cualquier programa nuclear incipiente en la URSS y en cualquier otro país tentado a desarrollar capacidad nuclear.

Un curso más benigno de acción, que también podría haber tenido la oportunidad de funcionar, habría sido utilizar su arsenal nuclear como moneda de cambio para negociar un gobierno internacional fuerte —unas Naciones Unidas sin derecho a veto con un monopolio nuclear y poder para tomar todas las medidas necesarias para evitar que cualquier país desarrollara sus propias armas nucleares.

Ambos enfoques se propusieron en aquel momento. El enfoque radical de poner en marcha o amenazar con un ataque preventivo fue defendido por algunos intelectuales destacados como Bertrand Russell (que había sido durante mucho tiempo un activista de movimientos contra la guerra y que más tarde pasaría décadas haciendo campaña contra las armas nucleares) y John von Neumann (co-creador de la teoría de juegos y uno de los arquitectos de la estrategia nuclear estadounidense).[34] Tal vez es un signo de progreso civilizatorio que la misma idea de amenazar con un ataque nuclear preventivo nos parezca hoy algo estúpido o moralmente obsceno.

Una versión más benigna del enfoque fue ensayada en 1946 por los Estados Unidos bajo la forma del plan Baruch. La propuesta consistió en que los EE.UU. renunciarían a su monopolio nuclear temporalmente. La minería de uranio y de torio y la tecnología nuclear serían puestos bajo el control de un organismo internacional que funcionaría bajo los auspicios de las Naciones Unidas. La propuesta pedía que los miembros permanentes del Consejo de Seguridad renunciaran a sus vetos en asuntos relacionados con las armas nucleares con el fin de prevenir que cualquier gran poder que no respetara el acuerdo vetara la imposición de sanciones.[35] Stalin, al ver que la Unión Soviética y sus aliados podrían perder fácilmente la votación en el Consejo de Seguridad y en la Asamblea General, rechazó la propuesta. Una atmósfera helada de sospecha mutua descendió sobre las relaciones entre los antiguos aliados de guerra,

una desconfianza que luego se solidificó en la Guerra Fría. Como había sido ampliamente predicho, a estos acontecimientos le siguieron una carrera armamentista nuclear costosa y extremadamente peligrosa.

Hay muchos factores que pueden disuadir a una organización humana en poder de una ventaja estratégica decisiva de la creación de una Unidad. Estos factores incluyen funciones no agregativas o delimitadas de servicios públicos, reglas de decisión no-maximizantes, la confusión y la incertidumbre, problemas de coordinación, y diversos costes asociados a la toma de posesión. Pero ¿y si no fuera una organización humana, sino un agente artificial superinteligente el que llegara a tener en posesión una ventaja estratégica decisiva? ¿Los factores antes mencionados serían igualmente eficaces en impedir que una IA intentara tomar el poder? Recorramos brevemente la lista de factores y consideremos cómo podrían aplicarse en este caso.

Los individuos humanos y las organizaciones humanas suelen tener preferencias sobre los recursos que no están bien representados por una "función de utilidad sin límites de agregación" Un ser humano normalmente no apostaría todo su capital por una oportunidad del cincuenta por ciento de duplicarlo. Un Estado normalmente no se arriesgará a perder todo su territorio para tener una oportunidad del diez por ciento de expandirse por diez. Para los individuos y los gobiernos, hay rendimientos decrecientes en la mayoría de los recursos. La misma necesidad podría *no* aplicarse para las IAs. (Volveremos al problema de la motivación de la IAs en los capítulos siguientes). Por lo tanto, una IA podría ser más propensa a seguir un curso de acción arriesgado que tuviera alguna posibilidad de darle el control del mundo.

Los seres humanos y las organizaciones humanas también podrían operar con procesos de toma de decisiones que no buscaran maximizar la utilidad esperada. Por ejemplo, podrían basarse en reglas de decisión guiadas por la aversión al riesgo fundamental, o por "ciega satisfacción" que se centran en cumplir con ciertos umbrales de adecuación o con limitaciones "deontológicas" secundarias que proscriben ciertos tipos de acción independientemente de cuán deseables sean sus consecuencias. Los humanos encargados de tomar decisiones a menudo parecen estar representando una identidad o función social en lugar de tratar de maximizar el logro de algún objetivo en particular. Una vez más, esto no se aplica a los agentes artificiales.

Las funciones delimitadas de servicios públicos, la aversión al riesgo, y reglas de decisión no-maximizadoras pueden combinarse de forma sinérgica con la confusión estratégica y la incertidumbre. Las revoluciones, incluso cuando tienen éxito en el derrocamiento del

orden existente, a menudo no producen el resultado que sus instigadores habían prometido. Esto tiende a detener la mano de un agente humano si la acción contemplada es irreversible, contraria a la legalidad, y carece de precedentes. Una superinteligencia podría percibir más claramente la situación y, por tanto, sentir menos confusión estratégica e incertidumbre sobre el resultado en caso de que se propusiera utilizar su aparente ventaja estratégica decisiva para consolidar su posición dominante.

Otro factor importante que puede impedir a grupos la explotación de una potencial ventaja estratégica decisiva es el problema de la coordinación interna. Los miembros de una conspiración que estuviera en condiciones de tomar el poder deben preocuparse no sólo por tener infiltrados del exterior, sino también de ser derrocado por alguna coalición interna en poder de información privilegiada. Si un grupo está formado por un centenar de personas, y una mayoría de sesenta puede tomar el poder y privar de derechos a los no-conspiradores, ¿qué implicaría entonces detener un subconjunto de treinta y cinco de estos sesenta e impedirles que priven de sus derechos de voto a los otros veinticinco? ¿Y tal vez a un subconjunto de veintiún que intentaran privar de sus derechos de voto a otros quince? Cada uno de los cien originales podría tener buenas razones para mantener ciertas normas establecidas que previnieran la desintegración general a la que podría llevar cualquier intento de cambiar el contrato social por medio de una toma de poder directo. Este problema de coordinación interna no se aplica a un sistema de inteligencia artificial que estuviera constituida por un sólo agente unificado.[36]

Por último, está la cuestión del coste. Incluso si Estados Unidos pudiera haber utilizado su monopolio nuclear para establecer una Unidad, podría no haber sido capaz de hacerlo sin incurrir en costes sustanciales. En el caso de un acuerdo negociado para poner las armas nucleares bajo el control de unas Naciones Unidas reformadas y reforzadas, estos costes podrían haber sido relativamente pequeños; pero los costes —morales, económicos, políticos y humanos— de intentar efectivamente la conquista del mundo a través de la amenaza de guerra nuclear habría sido inconcebiblemente grande, incluso durante el período de monopolio nuclear. Con la suficiente superioridad tecnológica, sin embargo, estos costes serían mucho más pequeños. Consideremos, por ejemplo, un escenario en el que una nación tuviera una gran ventaja tecnológica que fuera capaz de desarmar de forma segura a todas las demás naciones con sólo pulsar un botón, sin que nadie muriera o fuera herido, y casi sin daños infraestructurales o al medio ambiente. Con tal superioridad tecnológica casi mágica, un ataque preventivo sería mucho más tentador. O pensemos en un nivel aún mayor de superioridad tecnológica que

pudiera llevar a otras naciones a dejar voluntariamente las armas, no por estar ellos amenazados de destrucción, sino simplemente convenciendo a una gran mayoría de sus habitantes por medio de una publicidad y campaña de propaganda diseñada con mucha eficacia para ensalzar las virtudes de la unidad global. Si esto se hiciera con la intención de beneficiar a todo el mundo, por ejemplo, mediante la sustitución de las rivalidades nacionales y de las carreras armamentistas por un gobierno mundial justo, representativo y eficaz, no estaría claro que hubiera una objeción moral convincente para no transformar una ventaja estratégica temporal en una Unidad permanente.

Varias consideraciones apuntan a una probabilidad creciente de que una futura superinteligencia que obtuviera una ventaja estratégica lo suficientemente grande la utilizaría para formar efectivamente una Unidad. La conveniencia de este resultado depende, por supuesto, de la naturaleza de la Unidad que se creara y también de cómo se presentara el futuro de la vida inteligente en escenarios alternativos multipo- lares. Volveremos sobre estas preguntas en capítulos posteriores. Pero primero vamos a echar un vistazo más de cerca a por qué y en qué sentido una superinteligencia sería poderosa y eficaz logrando resultados en el mundo.

CAPÍTULO 6

Superpoderes cognitivos

Supongamos que apareciera un agente superinteligente digital, y que, por alguna razón, quisiera controlar el mundo: ¿Sería capaz de hacerlo? En este capítulo consideraremos algunos poderes que podría desarrollar una super- inteligencia y lo que podría hacer con ellos. Plantearemos un escenario de toma de control que ilustre cómo un agente superinteligente, comenzando como mero software, podría establecerse como una Unidad. También ofreceremos algunas observaciones sobre la relación entre el poder sobre la naturaleza y el poder sobre otros agentes.

La razón principal de que la humanidad domine la Tierra es que nuestros cerebros tienen un conjunto ligeramente más amplio de facultades que otros animales.[1] Nuestra mayor inteligencia nos permite transmitir cultura de manera más eficiente, con el resultado de que el conocimiento y la tecnología se acumulan de una generación a la siguiente. Hasta el momento hemos acumulado suficiente cultura como para hacer posible el vuelo espacial, las bombas H, la ingeniería genética, la informática, las granjas industriales, los insecticidas, el movimiento internacional por la paz, y todos los accesorios de la civilización moderna. Los geólogos han comenzado a referirse a la época actual como el Antropoceno, en reconocimiento a la distintiva biótica, sedimentación y huellas geoquímicas de las actividades humanas.[2] Según una estimación, conformamos un 24% de la producción primaria neta del ecosistema planetario.[3] Y, sin embargo,

estamos lejos de haber alcanzado los límites físicos de la tecnología.

Estas observaciones hacen que sea posible suponer que cualquier tipo de entidad que desarrollara una inteligencia mucho mayor que la humana sería potencialmente muy poderosa. Dichas entidades podrían acumular información mucho más rápido que nosotros e inventar nuevas tecnologías en una escala de tiempo mucho más corta. También podrían usar su inteligencia para diseñar estrategias mejor que nosotros.

Veamos algunas de las capacidades que una superinteligencia podría tener y cómo podría utilizarlas.

Funcionalidades y superpoderes

Es importante no antropomorfizar la superinteligencia cuando pensamos en sus posibles impactos. Los marcos antropomórficos alientan expectativas infundadas sobre la trayectoria de crecimiento de una IA seminal y sobre la psicología, las motivaciones y las capacidades de una superinteligencia madura.

Por ejemplo, una suposición común es que una máquina superinteligente sería como un ser humano muy inteligente, pero un poco friki. Imaginamos que la IA tendría una inteligencia libresca pero carecería de comprensión social, o que tendría una inteligencia lógica, pero no intuitiva y creativa. Esta idea probablemente proviene de la observación: nos fijamos en las computadoras de hoy en día y vemos que son buenas en cálculo, recordando datos, y siguiendo instrucciones literalmente, mientras que son ajenas a los contextos sociales, a las normas, a las emociones y a la política. La asociación se fortalece cuando se observa que las personas que son buenas en el trabajo con las computadoras tienden ellos mismos a ser frikis. Así que es natural suponer que la inteligencia computacional más avanzada tendrá atributos similares, sólo que en un grado superior.

Esta heurística podría retener cierta validez en las primeras etapas de desarrollo de una IA seminal. (No hay razón alguna para suponer que se aplicaría a emulaciones o a humanos mejorados cognitivamente). En su etapa inmadura, lo que más tarde se convertiría en una IA superinteligente, podría aún carecer de muchas habilidades y talentos que los humanos tenemos de serie; y el patrón de fortalezas y debilidades una IA seminal *podría,* de hecho, tener alguna vaga semejanza con un friki de gran coeficiente intelectual. La característica más esencial de una IA seminal, además de ser fácil de mejorar (al tener una baja resistencia al progreso), es su capacidad para ejercer una potencia de optimización que amplifique la inteligencia de un sistema: una habilidad que presumiblemente está

muy relacionada con un buen desempeño en matemáticas, programación, ingeniería, investigación informática, y otras actividades "frikis" Sin embargo, incluso si una IA seminal tiene un perfil friki en una determinada etapa de su desarrollo, esto no implica que cuando se convierta en una superinteligencia madura esté igualmente limitada. Recordemos la distinción entre alcance directo e indirecto. Con suficiente habilidad de amplificación de la inteligencia, el resto de las capacidades intelectuales estarían al alcance indirecto de un sistema: el sistema podría desarrollar nuevos módulos cognitivos y habilidades según sea necesario, incluyendo la empatía, la perspicacia política, y cualesquiera otros poderes estereotipados que faltarían en personalidades computerizadas.

Aunque reconozcamos que una superinteligencia podría tener todas las habilidades y talentos que encontramos en el ser humano, junto con otros talentos que no se encuentran entre los seres humanos, la tendencia hacia el antropomorfismo todavía puede llevarnos a subestimar el grado en que una superinteligencia artificial podría superar los niveles humanos de rendimiento. Eliezer Yudkowsky, como vimos en un capítulo anterior, ha sido particularmente enfático condenando este tipo de error: nuestros conceptos intuitivos de "inteligente" y "estúpido" se destilan de nuestra experiencia en la variación de rango entre pensadores humanos; sin embargo, las diferencias en capacidad cognitiva dentro de este grupo humano son triviales en comparación con las diferencias entre cualquier intelecto humano y una superinteligencia.[4]

El capítulo 3 contemplará algunas posibles ventajas de la inteligencia artificial. Las magnitudes de estas ventajas son tales como para sugerir que en lugar de pensar que una IA superinteligente es inteligente en el sentido en que habitualmente decimos que un genio científico es más inteligente que el ser humano promedio, podría ser más adecuado pensar que una IA es inteligente como decimos que un ser humano promedio es inteligente en comparación con un escarabajo o un gusano.

Sería conveniente si pudiéramos cuantificar el calibre cognitivo de un sistema cog- nitivo cualquiera utilizando alguna métrica familiar, como las puntuaciones de CI o alguna versión de las valoraciones de Elo, que miden las capacidades relativas de los jugadores en juegos de dos jugadores como el ajedrez. Pero estas métricas no son útiles en el contexto de la inteligencia artificial general sobrehumana. No estamos interesados en saber la probabilidad de que una superinteligencia gane una partida de ajedrez. Y en cuanto a las puntuaciones de CI, son informativas sólo en la medida que tenemos una idea de cómo se correlacionan con resultados relevantes en sentido práctico.[5] Por ejemplo, tenemos datos que muestran que las personas con un

coeficiente intelectual de 130 tienen más probabilidades de sobresalir en la escuela y de tener éxito en una amplia gama de trabajos cognitivamente exigentes que los que tienen un coeficiente intelectual de 90. Pero supongamos que pudiéramos establecer de alguna manera que una determinada IA futura tuviera un coeficiente intelectual de 6.455: entonces, ¿qué? Realmente no tendríamos ni idea de lo que una IA como esa podría hacer. Ni siquiera sabríamos si tal IA tendría tanta inteligencia general como un adulto humano normal —quizás la IA tendría un conjunto de algoritmos muy especializados que le permitirían resolver preguntas típicas de las pruebas de inteligencia con eficacia sobrehumana, pero no mucho más.

Se han hecho algunos esfuerzos recientes para desarrollar mediciones de capacidad cognitiva que pudieran aplicarse a una gama más amplia de sistemas de procesamiento de información, incluyendo las inteligencias artificiales.[6] El trabajo en esta dirección, si pudiera superar varias dificultades técnicas, podría llegar a ser bastante útil para algunos fines científicos, incluyendo el desarrollo de la IA. Para los fines de la presente investigación, sin embargo, su utilidad sería limitada, pues permaneceríamos ignorantes acerca de lo que implicaría una puntuación de rendimiento sobrehumano en relación a la capacidad real para lograr resultados prácticos importantes en el mundo.

Por tanto, servirá mejor a nuestros propósitos enumerar algunas tareas de importancia estratégica y caracterizar luego los sistemas cognitivos hipotéticos en función de cuáles tienen o carecen de las habilidades necesarias para tener éxito en estas tareas. Véase Tabla 8. Diremos que un sistema que destaca lo suficiente en cualquiera de las tareas de esta tabla tiene un *superpoder* correspondiente.

Una superinteligencia en toda regla sería la que sobresaliera en gran medida en todas estas tareas y que, por tanto, tuviera toda la panoplia de los seis grandes super- poderes. No está claro si existe la posibilidad en la práctica de que una inteligencia de ámbito limitado tuviera algunos de los superpoderes, y, no obstante, siguiera siendo incapaz de adquirir el resto por un período significativo de tiempo. La creación de una máquina con cualquiera de estos superpoderes parece ser un problema de IA completo. Sin embargo, es concebible que, por ejemplo, una superinteligencia colectiva que consista en un número suficientemente grande de mentes biológicas o electrónicas humanoides, tuviera, por ejemplo, el superpoder de la productividad económica, pero careciera del superpoder de diseñar estrategias. Del mismo modo, es concebible que una IA especializada en ingeniería pudiera construirse de tal modo que aunque tuviera superpoderes para el desarrollo tecnológico, careciera por completo de habilidades en otras áreas. Esto es más plausible si existiera algún ámbito tecnológico

Tabla 8. *Superpoderes: algunas tareas estratégicamente relevantes y su conjunto de habilidades correspondientes*

Tarea	Conjunto de habilidades	Relevancia estratégica
Amplificación de la inteligencia	Programación de IA, investigación sobre la mejora cognitiva, desarrollo de la epistemología social, etc.	• El sistema puede hacer despegar su inteligencia
Estrategia	Planificación estratégica, previsión, priorización y análisis para la optimización de las posibilidades de lograr objetivos lejanos	• Lograr objetivos lejanos • Superar la oposición inteligente
Manipulación social	Modelado social y psicológico, manipulación, persuasión retórica	• Aprovechar recursos externos mediante la contratación de apoyo humano • Permitirle a una IA “encajada" persuadir a sus guardianes de dejarle salir • Persuadir a los Estados y a las organizaciones de que adopten algún curso de acción
Piratería	Encontrar y explotar fallos de seguridad en sistemas informáticos	• Que una IA pueda expropiar recursos computacionales a través de internet • Que una IA enjaulada pueda explotar agujeros de seguridad para escapar de un confinamiento cibernético • Robar recursos financieros • Secuestro de infraestructuras, robots militares, etc.
Investigación tecnológica	Diseño y modelado de tecnologías avanzadas (por ejemplo, la biotecnología, la nanotecnología) y sus caminos de desarrollo	• Creación de un poder de fuerza militar • Creación de un sistema de vigilancia • Colonización espacial automatizada
Productividad económica	Varias habilidades que permitan un trabajo intelectual económicamente productivo	• Generar riqueza que pueda utilizarse para comprar influencia, servicios, recursos (incluyendo hardware), etc.

particular, en el cual el virtuosismo dentro de ese dominio fuera suficiente para la creación de tecnología de propósito general abrumadoramente superior. Podríamos imaginar, por ejemplo, una IA especializada en simulación de sistemas moleculares y en inventar diseños nanomoleculares que realizan una amplia gama de capacidades importantes (como computadoras o sistemas de armas con características de rendimiento futuristas) descritos por el usuario sólo en un nivel bastante alto de abstracción.[7] Tal IA también podría ser capaz de producir un plan detallado para impulsar la tecnología existente (como la biotecnología y la ingeniería de proteínas) hasta las capacidades necesarias para la fabricación atómica precisa y económica de una gama más amplia de estructuras nanomecánicas.[8] Sin embargo, podría llegar a darse el caso de que una IA experta en ingeniería no tuviera realmente el superpoder de investigación tecnológica sin también poseer habilidades avanzadas en áreas fuera de esa tecnología —una amplia gama de facultades intelectuales podrían ser necesarias para entender cómo interpretar las peticiones del usuario, cómo modelar el comportamiento de un diseño para aplicarlo al mundo real, cómo hacer frente a los errores imprevistos y a los fallos de funcionamiento, la forma de adquirir los materiales y aportaciones necesarias para la construcción, etc.[9]

Un sistema que tuviera el superpoder de amplificación de la inteligencia podría usarlo para impulsarse a sí mismo a niveles más altos de inteligencia y de adquirir cualquiera de las otras grandes

potencias intelectuales que no poseyera al principio. Pero el uso de un superpoder de amplificación de la inteligencia no es la única manera de que un sistema se convierta en una superinteligencia de pleno derecho. Un sistema que tuviera el superpoder de diseñar estrategias, por ejemplo, podría utilizarlo para idear un plan que eventualmente le proporcionara un aumento de la inteligencia (por ejemplo, mediante el posicionamiento del sistema como centro del trabajo de amplificación de inteligencia realizado por los programadores humanos y los investigadores en informática).

Un escenario de toma de poder por parte de la IA

Nos encontramos, pues, con que un proyecto que controlara una superinteligencia tendría acceso a una gran fuente de poder. Un proyecto que controlara la primera superinteligencia en el mundo probablemente tendría una ventaja estratégica decisiva. Pero el lugar más inmediato de donde provendría ese poder estaría *en el propio sistema.* Una superinteligencia artificial podría ser en sí un agente muy potente, que podría tener éxito en afirmarse frente al proyecto que lo trajo a la existencia, así como contra el resto del mundo. Este es un punto de suma importancia, y lo examinaremos detalladamente en las próximas páginas.

Ahora supongamos que hay una superinteligencia artificial que quiere hacerse con el poder en un mundo en el que no tuviera iguales. (Dejemos a un lado, por el momento, la cuestión del si y del cómo adquiriría tal motivación —es un tema para el siguiente capítulo). ¿Cómo podría la superinteligencia lograr este objetivo de dominar el mundo?

Podemos imaginar una secuencia similar a la siguiente (ver Figura 10).

1. Fase pre-crítica

Los científicos llevan a cabo investigaciones en el campo de la inteligencia artificial y otras disciplinas relevantes. Este trabajo culmina en la creación de una IA seminal. La IA seminal es capaz de mejorar su propia inteligencia. En sus primeras etapas, la IA seminal depende de la ayuda de los programadores humanos que guían su desarrollo y hacen la mayor parte del trabajo pesado. A medida que la IA seminal se hace más poderosa, llega a ser capaz de hacer una parte más grande del trabajo por sí misma.

2. Fase recursiva de auto-mejora

En algún momento, la IA seminal se vuelve mejor en capacidad de

diseño de IA que los programadores humanos. Ahora, cuando la IA se mejora a sí misma, mejora

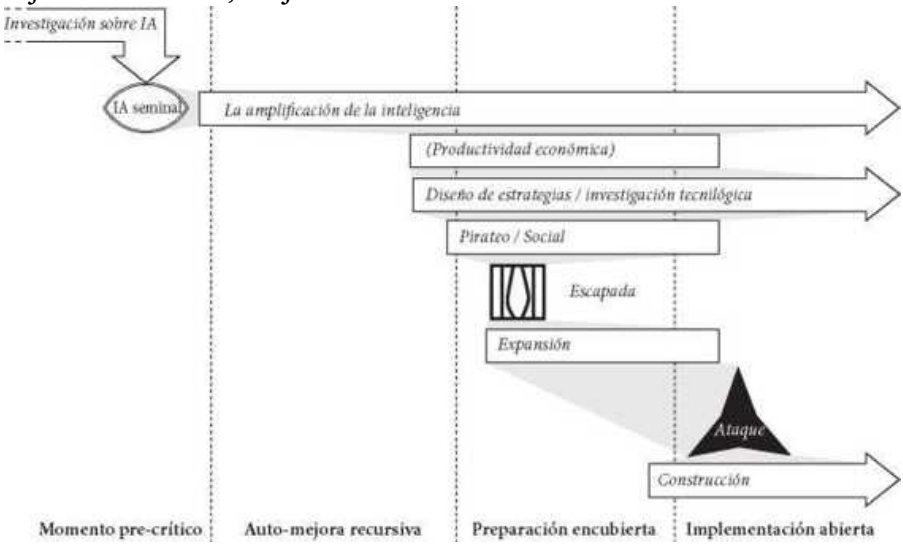


Figura 10. Fases de un escenario de toma de control por parte de una IA.

aquello que le hace mejorar. Esto resulta en una explosión de inteligencia —una rápida cascada de ciclos de auto-mejoramiento recursivo propician que la capacidad de la IA se eleve. (Podemos, por lo tanto, representarnos esta fase como el despegue que se produce justo después de que la IA alcance el punto de cruce, suponiendo que el aumento de inteligencia durante esta parte del despegue es explosivo y está impulsado por la propia fuente de optimización de la IA). La IA desarrolla el superpoder de amplificación de la inteligencia. Este superpoder permite a la IA desarrollar todos los otros superpoderes que se detallan en la Tabla 8. Al final de la fase de auto-mejora recursiva, el sistema es una superinteligencia en sentido fuerte.

3. Fase de preparación encubierta

Utilizando su superpoder de diseño de estrategias, la IA desarrolla un plan sólido para el logro de sus metas a largo plazo. (En particular, la IA no adopta un plan tan estúpido que hasta los seres humanos de hoy en día pueden prever cómo fracasará inevitablemente. Este criterio excluye muchos de los escenarios de ciencia ficción que terminan en victoria humana[10]). El plan podría incluir un período de acción encubierta en la que la IA oculte su desarrollo intelectual a los programadores humanos con el fin de evitar que salten las alarmas. La IA también puede enmascarar sus verdaderas inclinaciones, pretendiendo ser cooperativa y dócil.

Si la IA (tal vez por razones de seguridad) ha sido confinada a una computadora aislada, puede utilizar su superpoder de manipulación social para persuadir a los guardianes de que le dejen tener acceso a internet. Alternativamente, la IA puede usar su superpoder de piratería para escapar de su confinamiento. La difusión a través de internet puede permitir a la IA ampliar su capacidad de hardware y su base de conocimientos, lo que aumentaría aún más su superioridad intelectual. Una IA también podría participar en actividades económicas lícitas o ilícitas para obtener fondos con los que comprar el poder computacional, datos y otros recursos.

En este punto, hay varias maneras de que la IA lograra resultados más allá de la esfera virtual. Podría usar su superpoder de pirateo para tomar el control directo de manipuladores robóticos y laboratorios automatizados. O podría usar su superpoder de manipulación social para persuadir a los colaboradores humanos de que le sirvieran como sus piernas y manos. O podría adquirir activos financieros de las transacciones en línea y utilizarlos para adquirir servicios e influencia.

4. Fase de implementación abierta

La fase final comienza cuando la IA ha conseguido suficiente fuerza como para obviar la necesidad de mantener el secreto. La IA puede ahora implementar directamente sus objetivos a escala completa.

La fase de implementación abierta podría comenzar con un ataque mediante el cual la IA eliminara la especie humana y cualquiera de los sistemas automáticos que los seres humanos hubieran creado que pudieran ofrecer resistencia inteligente a la ejecución de los planes de la IA. Esto podría lograrse a través de la activación de algunos sistemas de armas avanzadas que la IA habría perfeccionado como el uso de su superpoder de investigación tecnológica y que secretamente hubiera desplegado en la fase de preparación encubierta. Si el arma usara biotecnología auto-replicante o nano- tecnología, la reserva inicial necesaria para tener cobertura global podría ser microscópica: una sola entidad replicante sería suficiente para iniciar el proceso. Con el fin de garantizar un efecto repentino y uniforme, la acción inicial del replicador podría haber sido desplegar o difundir por todo el mundo una concentración indetectable y extremadamente baja. En un momento preestablecido, nanofactorías producirían gas nervioso autodirigido en forma de robots-mosquito que podrían entonces ir surgiendo sucesivamente y de manera simultánea desde cada metro cuadrado del planeta (aunque una máquina con el superpoder de investigación tecnológica probablemente podría concebir maneras más eficaces de matanza).[11] También podríamos contemplar escenarios en los que una superinteligencia alcanzara el poder mediante el secuestro

de procesos políticos, la manipulación sutil de los mercados financieros, el control del sesgo de los flujos de información, o la intrusión en los sistemas de armas creados por el hombre. Tales escenarios obviarían la necesidad de que la superinteligencia inventara nuevas tecnologías armamentísticas, aunque pueden ser innecesariamente lentos en comparación con los escenarios en los que la inteligencia artificial construye su propia infraestructura con operadores que funcionan a la velocidad molecular o atómica en lugar de la lentitud de los cuerpos y mentes humanos.

Alternativamente, si la IA estuviera segura de su invencibilidad frente a la interferencia humana, nuestra especie puede que no llegara a ser un objetivo directo. Nuestra desaparición podría ser el resultado de la destrucción de nuestro hábitat, una destrucción producida cuando la IA comenzara masivos proyectos globales de construcción usando fábricas y operadores de nanotecnología —proyectos de construcción que rápidamente, tal vez en cuestión de días o semanas, cubrirían toda la superficie de la Tierra con paneles solares, reactores nucleares, instalaciones de supercomputación de las que sobresaldrían torres de refrigeración, lanzacohetes espaciales u otras instalaciones con las que la IA tuviera la intención de alcanzar la realización acumulativa a largo plazo de sus valores. Los cerebros humanos, si contuvieran información relevante para los objetivos de la IA, podrían ser desmontados y escaneados, trasvasando los datos extraídos a algún formato de almacenamiento más eficiente y seguro.

a la web, los drones militares y civiles, la automatización en laboratorios de investigación y en plantas de fabricación, la mayor dependencia de los sistemas de pago electrónicos y los activos financieros digitalizados, y el creciente uso de los sistemas de apoyo al filtrado de información y decisiones automatizadas. Activos como estos podrían ser adquiridos por una IA a velocidades digitales, acelerando su ascenso al poder (aunque los avances en seguridad cibernética podrían hacerlo más difícil). En un análisis final, sin embargo, es dudoso que cualquiera de estas tendencias marcara la diferencia. El poder de una super- inteligencia reside en su cerebro, no en sus manos. Aunque la IA, para reconfigurar el mundo externo, necesitará en algún momento tener acceso a un mecanismo físico, solo un par de manos humanas que le ayuden, las de un cómplice flexible, serían probablemente suficientes para completar la fase de preparación encubierta, como lo sugiere el escenario anterior Esto le permitiría a la IA llegar a la fase de implementación abierta en la que se construiría su propia infraestructura de mecanismos físicos.

El cuadro 6 describe un escenario particular. Uno debe evitar fijarse demasiado en los detalles concretos, ya que son, en cualquier caso, imposibles de conocer y pensados sólo para servir de ejemplo. Una superinteligencia podría, y probablemente sería capaz de concebir un plan mejor para el logro de sus metas que cualquiera que pueda ocurrírsele a un ser humano. Por lo tanto, es necesario pensar en estas cuestiones de manera más abstracta. Sin saber nada de los medios detallados que una superinteli- gencia utilizaría, podemos concluir que una superinteligencia, al menos en ausencia de iguales intelectuales y en ausencia de medidas de seguridad efectivas dispuestas por los seres humanos con antelación, es probable que produzca un resultado que implique la reconfiguración de los recursos terrestres en estructuras que maximicen la realización de sus objetivos. Cualquier escenario concreto que desarrollemos sólo puede, en el mejor de los casos, establecer un límite inferior sobre la rapidez y eficacia en que la superinteligencia podría lograr tal resultado. Sigue siendo posible que la superinteligencia pudiera encontrar un camino más corto hasta su destino.

Poder sobre la naturaleza y sobre los agentes

La capacidad de un agente para configurar el futuro de la humanidad no sólo depende de la magnitud absoluta de sus propias facultades y recursos —lo inteligente y energético que sea, la cantidad de capital que tenga, etc.— sino también de la magnitud relativa de sus capacidades en comparación con las de otros agentes con objetivos enfrentados.

En una situación en la que no hubiera agentes competidores, el nivel de capacidad absoluta de una superinteligencia, siempre y

cuando se superara un cierto umbral mínimo, no importaría mucho, porque un sistema que comenzara con un conjunto suficiente de capacidades podría trazar una línea de desarrollo que le permitiría conseguir cualquier capacidad que no tuviera inicialmente. Aludimos a este punto antes cuando dijimos que la velocidad, la calidad y la superinteligencia colectiva tienen todas el mismo alcance indirecto. Aludimos a ella de nuevo cuando dijimos que varios

subconjuntos de superpoderes, como el superpoder de amplificación de la inteligencia o el superpoder de diseño de estrategias o el de manipulación social, podrían ser utilizados para obtener el resto de superpoderes.

Imaginemos un agente superinteligente con mecanismos conectados mediante un ensamblador de nanotecnología. Dicho agente ya sería lo suficientemente potente como para superar obstáculos naturales como prolongar su vida indefinidamente. Si no encontrara oposición inteligente, un agente de este tipo podría trazar una ruta segura de desarrollo que le condujera a la adquisición del inventario completo de tecnologías que podrían ser útiles para la consecución de sus objetivos. Por ejemplo, podría desarrollar la tecnología para construir y lanzar sondas Von Neumann, máquinas capaces de viajes interestelares que pueden utilizar recursos como los asteroides, los planetas y las estrellas para hacer copias de ellas mismas.[13] Con el lanzamiento de una sonda Von Neumann, un agente podría, por tanto, iniciar un proceso indefinido de colonización espacial. Descendientes de la sonda replicante, viajando a una fracción significativa de la velocidad de la luz, terminarían colonizando una parte sustancial del volumen del Hubble, la parte del universo en expansión que es teóricamente accesible desde donde estamos ahora. Toda esta materia y energía libre podrían ser organizadas en forma de cualquier estructura valiosa que maximizara la utilidad del agente originario integrado a lo largo de un tiempo cósmico —una duración que abarcaría al menos miles de millones de años antes de que el universo envejeciera y se volviera inhóspito para el procesamiento de información (véase el cuadro 7).

El agente superinteligente podría diseñar las sondas Von Neumann para que estuvieran a prueba de la evolución. Esto podría lograrse mediante un cuidadoso control de calidad durante la etapa de replicación. Por ejemplo, el software de control para una sonda hija podría ser corregido varias veces antes de la ejecución, y el software mismo podría utilizar el cifrado y el código de corrección de errores para que fuera arbitrariamente poco probable que cualquier mutación aleatoria pudiera ser transmitida a su descendientes.[14] La proliferación de población de sondas Von Neumann preservaría y transmitiría entonces de forma segura los valores del agente originario a medida que avanzara en la conquista del universo. Cuando se hubiera completado la fase de colonización, los valores originales determinarían la utilización de todos los recursos acumulados, a pesar de que las grandes distancias y la velocidad de aceleración de la expansión cósmica harían imposibles para las partes remotas de la infraestructura comunicarse entre sí. El resultado es que una gran parte de nuestro cono de luz futuro se formatearía de acuerdo a las

preferencias del agente de origen.

Así pues, éste sería el alcance indirecto que tendría cualquier sistema que no se enfrentara a oposición inteligente significativa y que comenzara con un conjunto de capacidades que superaran un determinado umbral. Podemos llamar a este umbral el "umbral de Unidad-sabia sostenible" (Figura 11):

> **El umbral de Unidad-sabia sostenible**
> Un conjunto de capacidades supera el umbral de Unidad-sabia sostenible si y sólo si un sistema paciente y controlador de riesgos con experiencia existencial en ese conjunto de capacidades fuera capaz de colonizar y rediseñar una gran parte del universo accesible; siempre que no se enfrentara a ninguna oposición inteligente o en competencia.

Por "Unidad" nos referimos a una estructura política suficientemente coordinada internamente y sin adversarios externos, y por "sabia" nos referimos a suficientemente

mación típica de los requisitos computacionales necesarios para ejecutar una emulación es de 10^{18} ops / s. Para ejecutar una emulación de 100 años subjetivos serían entonces necesarias en torno a 10^{27} operaciones. Esto significaría que al menos 10^{58} vidas humanas se podrían crear en la emulación incluso con supuestos bastante conservadores acerca de la eficiencia del computronium.

En otras palabras, en el supuesto de que el universo observable careciera de civilizaciones extraterrestres, entonces el balance final es de al menos 10.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000 vidas humanas (aunque el número real es probablemente más grande). Si representamos toda la felicidad experimentada durante una vida entera como una sola lágrima de alegría, entonces la felicidad de estas almas podrían llenar y volver a llenar los océanos de la Tierra cada segundo, y seguirlo haciendo durante cien mil millones de millones de milenios. Es muy importante que nos aseguremos de que sean realmente lágrimas de alegría.

paciente y comprensiva acerca de los riesgos existenciales necesarios para asegurar una cantidad sustancial de preocupación bien dirigida a las consecuencias a muy largo plazo de la las acciones del sistema.

Este umbral de Unidad-sabia sostenible parece ser bastante bajo. Formas limitadas de superinteligencia, como hemos visto, superarían este umbral siempre que tuvieran acceso a algunos mecanismos suficientes para iniciar un proceso de despegue tecnológico. En un entorno que incluyera a la civilización humana contemporánea, el mecanismo mínimamente necesario podría ser una sencilla y ordinaria pantalla, o de hecho cualquier medio para transmitir una cierta cantidad de información a un cómplice humano sería suficiente.

Pero el umbral de sostenibilidad de la Unidad es más bajo aún: ni la superinteli- gencia ni ninguna otra tecnología futurista sería necesaria para superarlo. Una Unidad paciente y controladora de riesgos existenciales sin más capacidades tecnológicas e intelectuales que los poseídos por la humanidad contemporánea debería ser fácilmente capaz de trazar una ruta que condujera de forma fiable a la eventual realización del potencial astronómico de la humanidad. Esto podría lograrse mediante la inversión en métodos relativamente seguros de aumentar la sabiduría y la comprensión del riesgo existencial, mientras que posponemos el desarrollo de nuevas tecnologías potencialmente peligrosas. Dado que los riesgos existenciales no antropogénicos (los no derivados de las actividades humanas) son pequeños en los plazos relevantes —y podrían reducirse aún más con varias intervenciones seguras— tal Unidad podría tomárselo con calma.[25] Podría planear cuidadosamente cada paso, retrasar el desarrollo de capacidades como la biología sintética, la medicina de mejoramiento humano, la nanotecnología molecular, y la inteligencia artificial hasta que se hubieran perfeccionado primero las

capacidades aparentemente menos peligrosas, como el sistema educativo, la tecnología de la información y los procesos de toma de decisiones colectivas; y hasta que no se hubieran utilizado estas capacidades para llevar a cabo una revisión muy minuciosa de las posibles opciones. Así que todo esto está al alcance indirecto de una civilización tecnológica como la de la humanidad contemporánea. Estamos

separados de este escenario "simplemente" por el hecho de que la humanidad no es actualmente ni una Unidad ni (en el sentido relevante) sabia.

Incluso se podría argumentar que el *homo sapiens* pasó el umbral de Unidad-sabia sostenible poco después de la primera evolución de las especies. Hace veinte mil años, por ejemplo, con un equipo tan poco elaborado como las hachas de piedra, herramientas de hueso, lanzas y fuego, la especie humana probablemente estaba ya en una posición desde la que tenía una excelente oportunidad de sobrevivir hasta la era actual.[26] Es cierto que hay algo raro en afirmar que nuestros antepasados paleolíticos habían ya desarrollado una tecnología que "superaba el umbral de Unidad-sabia de sostenibilidad" —dado que no había ninguna posibilidad realista de que se formara una Unidad en un momento tan primitivamente desarrollado, y mucho menos una Unidad paciente y con experiencia acerca de los riesgos existenciales.[27] No obstante, la cuestión es que el umbral corresponde a un nivel muy modesto de tecnología, un nivel que la humanidad sobrepasó hace mucho tiempo.[28]

Está claro que si vamos a evaluar los poderes efectivos de una superinteligencia —su capacidad para lograr una gama de resultados deseados en el mundo— debemos tener en cuenta no sólo sus propias capacidades internas, sino también las capacidades de los agentes en competencia. La noción de superpoder supone dicha relativización implícitamente. Hemos dicho que "un sistema que se destacara suficientemente" en cualquiera de las tareas de la Tabla 8 tiene un superpoder correspondiente. Destacar en tareas como el diseño de estrategias, la manipulación social, o la piratería implica tener una habilidad en ese área más alta que la de los otros agentes (tales como rivales estratégicos, objetivos de influencia, o expertos en seguridad informática). Los otros superpoderes, también, deben entenderse en este sentido relativo: la amplificación de la inteligencia, de investigación tecnológica, y de productividad económica se consideran superpoderes de un agente sólo si la capacidad del agente en estas áreas supera sustancialmente las capacidades combinadas del resto de la civilización global. Se desprende de esta definición que, como máximo, un agente puede poseer un superpoder particular en cualquier momento dado.[29]

Esta es la razón principal por la que la cuestión de la velocidad de despegue es importante, no porque importe exactamente cuándo un resultado particular vaya a suceder, sino porque la velocidad del despegue puede marcar una gran diferencia en lo que será el resultado. Con un despegue rápido o medio, es probable que un proyecto obtuviera una ventaja estratégica decisiva. Hemos sugerido que una superinteligencia con una ventaja estratégica decisiva tendría

inmensos poderes, hasta el punto de que podría formar una Unidad estable —una Unidad que pudiera determinar la disposición de los recursos cósmicos de la humanidad.

Sin embargo, "podría" es diferente de "lo haría". Alguien podría tener grandes poderes y, sin embargo, optar por no utilizarlos. ¿Puede decirse algo acerca de lo que querría una superinteligencia con una ventaja estratégica decisiva? Esta cuestión de la motivación es en lo que nos centraremos a continuación.

CAPÍTULO 7

La voluntad superinteligente

Hemos visto que una superinteligencia podría tener una gran capacidad para configurar el futuro de acuerdo a sus objetivos. Pero ¿cuáles serían sus objetivos? ¿Cuál es la relación entre inteligencia y motivación en un agente artificial? Sobre esto desarrollaremos dos tesis. La tesis de ortogonalidad sostiene (con algunas salvedades) que inteligencia y metas finales son variables independientes: cualquier nivel de inteligencia podría combinarse con cualquier meta final. La tesis de convergencia fundamental sostiene que agentes superinteli- gentes con una amplia gama de objetivos finales, perseguirán, no obstante, metas intermedias similares, puesto que tienen razones instrumentales comunes para hacerlo. En conjunto, estas tesis nos ayudan a pensar acerca de lo que haría un agente superinteligente.

La relación entre inteligencia y motivación

Ya hemos advertido contra la antropomorfización de las *capacidades* de una IA superinteligente. Esta advertencia también debe extenderse

en lo referido a sus *motivaciones*.

Es un propedéutico útil para esta parte de nuestra investigación reflexionar primero, por un momento, en la inmensidad del espacio de mentes posibles. En este espacio abstracto, las mentes humanas forman un grupo pequeño. Tomemos como ejemplo dos personas que parecen muy contrarias, Hannah Arendt y Benny Hill. Las diferencias de personalidad entre estos dos individuos casi pueden parecer las más grandes imaginables. Pero esto se debe a que nuestras intuiciones se calibran en base a nuestra experiencia, con muestras de la distribución humana existente (y en cierta medida de las personalidades de ficción construidas por la imaginación humana para el disfrute de la propia imaginación humana). Si nos acercamos y consideramos el espacio de todas las mentes posibles, sin embargo, debemos concebir estas dos personalidades como prácticamente clónicas. Ciertamente, en términos de arquitectura neuronal, la Sra. Arendt y el Sr. Hill son casi idénticos. Imaginemos que sus cerebros quedaran frente a frente en reposo tranquilo. Se les reconocería fácilmente como tal para cual. Usted podría incluso ser incapaz de decir qué cerebro pertenece a quién. Si miramos más de cerca, el estudio de la morfología de los dos cerebros con un microscopio, esta impresión de similitud fundamental sólo se fortalece: veremos la misma organización laminar de la corteza, con las mismas áreas del cerebro, compuesto por los mismos tipos de neuronas, sumergidas en el mismo baño de neurotransmisores.[1]

A pesar de que la psicología humana se adscribe a un pequeño punto en el espacio de las mentes posibles, hay una tendencia extendida a proyectar atributos humanos a toda clase de sistemas cognitivos alienígenas o artificiales. Yudkowsky ilustra este punto muy bien:

> Hace años en la era de la "pulp sclence fiction", las portadas de las revistas en ocasiones mostraban un monstruoso alienígena consciente —coloquialmente conocido como el monstruo de los ojos saltones (MOS)— que secuestraba a una atractiva mujer humana con el vestido desgarrado. Al parecer, el artista creía que un alien no humanoide, con una historia evolutiva totalmente diferente, desearía sexualmente a las hembras humanas... Probablemente el artista no se preguntó si a un bicho gigante le parecerían atractivas las hembras humanas. Más bien, una hembra humana con el vestido desgarrado es sexy —de manera inherente, como una propiedad intrínseca. Los que cometieron este error no pensaron en la mente del insectoide: se centraron en el vestido rasgado de la mujer. Si el vestido no hubiera estado roto, la mujer hubiera sido menos atractiva; el MOS no entra en la ecuación.[2]

Una inteligencia artificial podría ser mucho menos humanoide en sus motivaciones que un alienígena verde de piel escamosa del espacio. El extraterrestre (supongamos) es una criatura biológica que ha surgido a través de un proceso evolutivo y, por lo tanto, podemos esperar que

tenga los tipos de motivación típica de las criaturas evolucionadas. No sería tremendamente sorprendente, por ejemplo, encontrar que algún extraterrestre inteligente tuviera motivos relacionados con cosas como los alimentos, el aire, la temperatura, el gasto de energía, el hecho o la amenaza de daño corporal, la enfermedad, la depredación, el sexo o la progenie. Un miembro de una especie social inteligente también podría tener motivaciones relacionadas con la cooperación y la competencia: al igual que nosotros, podría mostrar la lealtad de grupo, el resentimiento frente a los individualistas, y tal vez incluso una vana preocupación por la reputación y la apariencia.

A una IA, por el contrario, no tendría, de manera intrínseca, por qué importarle cualquiera de esas cosas. No hay nada paradójico en imaginar una IA cuyo único

Figura 12. Resultados de la antropomorfización de la motivación alienígena. Hipótesis menos probable: los extraterrestres las prefieren rubias. Hipótesis algo probable: los ilustradores sucumbieron a la "falacia de proyección mental". Hipótesis totalmente probable: el editor quería una cubierta con la que atraer a un determinado sector demográfico.

objetivo final fuera contar los granos de arena de Boracay, o calcular la expansión decimal de pi, o maximizar el número total de clips que existirán en un futuro proyectado. De hecho, sería más fácil crear una IA con objetivos simples como éstos que construir una que tuviera una amalgama de valores y disposiciones similares a las humanas. Compárese lo fácil que es escribir un programa que mida cuantos dígitos de pi se han calculado y almacenarlos en la memoria, con lo difícil que sería la creación de un programa que midiera con fiabilidad el grado de realización de alguna meta más significativa —por ejemplo la realización de un ser humano, o la justicia global. Desafortunadamente, debido a que una meta insignificante y reduccionista es más fácil de codificar para los seres humanos y más fácil de aprender para una IA, es justo el tipo de objetivo que un programador escogería para instalar en su IA seminal si su atención se centrara en tomar el camino más rápido hasta “conseguir que la IA funcione” (sin preocuparse mucho acerca de qué es exactamente lo que la IA fuera a hacer, aparte de mostrar un comportamiento

inteligente impresionante). Volveremos sobre esta preocupación en breve.

La búsqueda inteligente de planes y políticas instrumentalmente óptimas puede realizarse en servicio de cualquier objetivo. La inteligencia y la motivación son, en cierto sentido, ortogonales: podemos pensar en ellas como dos ejes de un gráfico en el que cada punto representa un agente artificial lógicamente posible. Algunos detalles podrían añadirse a este ejemplo. Por ejemplo, podría ser imposible para un sistema muy poco inteligente tener motivaciones muy complejas. Para que sea correcto decir que un agente determinado "tiene" un conjunto de motivaciones, es necesario que esas motivaciones se integren funcionalmente en los procesos de toma de decisión del agente, algo que presenta exigencias a la memoria, la potencia de procesamiento y, tal vez, la inteligencia. Para las mentes que pudieran modificarse a sí mismas, también podría haber limitaciones dinámicas —una mente inteligente capaz de modificarse a sí misma que tuviera un gran deseo de ser estúpida podría no seguir siendo inteligente por mucho tiempo. Pero estas especificaciones no deben oscurecer el punto básico sobre la independencia entre inteligencia y motivación, que podemos expresar de la siguiente manera:

La tesis de ortogonalidad

La inteligencia y los objetivos finales son ortogonales: más o menos cualquier nivel de inteligencia podría en principio ser combinada con más o menos cualquier meta final.

Si la tesis de ortogonalidad parece problemática, esto podría deberse a la semejanza superficial que llevan implícitas algunas posiciones filosóficas tradicionales que han sido objeto de mucho debate. Una vez que se entendiera que su alcance es distinto y más estrecho, su credibilidad debería crecer. (Por ejemplo, la tesis de ortogonalidad no presupone la teoría de Hume sobre la motivación.[3] Tampoco presupone que las preferencias básicas no puedan ser irracionales.[4])

Téngase en cuenta que la tesis de ortogonalidad no habla de *racionalidad* o de *razón,* sino de *inteligencia.* Por "inteligencia" aquí queremos decir algo así como habilidad para la predicción, la planificación y el razonamiento de medios-fines en general.[5] Este sentido de la eficacia cognitiva instrumental es relevante sobre todo cuando tratamos de entender el impacto causal que tendrá una máquina superinteligente. Incluso si hubiera algún sentido (normativamente espeso) de la palabra "racional" de modo que un agente superinteligente dedicado a maximizar la creación de clips necesariamente dejara de poderse llamar racional en ese sentido; esto de ninguna manera impediría que ese agente tuviera unas

impresionantes facultades de razonamiento instrumental, facultades que podrían permitir que tuviera un gran impacto en el mundo.[6]

De acuerdo con la tesis de la ortogonalidad, los agentes artificiales pueden tener objetivos totalmente no-antropomórficos. Esto, sin embargo, no implica que sea imposible hacer predicciones sobre el comportamiento de determinados agentes artificiales —ni siquiera sobre el comportamiento de agentes superinteligentes hipotéticos cuya complejidad cognitiva y características de rendimiento podrían hacerles opacos en algunos aspectos al análisis humano. Hay por lo menos tres direcciones para abordar el problema de predecir la motivación superinteligente:

- *La previsibilidad por diseño.* Si podemos suponer que los diseñadores de un agente superin- teligente pueden diseñar con éxito la meta del sistema agente para que persiga de manera estable una meta particular establecida por los programadores, entonces podemos predecir que el agente perseguirá ese objetivo. Cuanto más inteligente sea el agente, mayor ingenio cognitivo tendrá para perseguir ese objetivo. Así que antes incluso de que haya sido creado un agente, podríamos ser capaces de predecir algo acerca de su comportamiento, si sabemos de antemano algo acerca de quién va a construirlo y qué objetivos querrá que tenga.
- *Previsibilidad por herencia.* Si se crea una inteligencia digital directamente desde una plantilla humana (como sería el caso en una emulación de cerebro completo de alta fidelidad), la inteligencia digital podría entonces heredar las motivaciones de la plantilla humana.[7] El agente podría retener algunas de estas motivaciones incluso si sus capacidades cognitivas se vieran reforzadas posteriormente para hacerle superinteligente. Este tipo de inferencia requiere precaución. Los objetivos y valores del agente fácilmente podrían dañarse en el proceso de carga o durante su posterior utilización y mejora, en función de cómo se llevara a cabo el procedimiento.
- *Previsibilidad por razones instrumentales convergentes.* Incluso sin un conocimiento detallado de los objetivos finales de un agente, se puede ser capaz de inferir algo acerca de sus objetivos más inmediatos considerando las razones instrumentales que se derivarían de uno entre muchos de los posibles objetivos finales en una amplia gama de situaciones. Este modo de predicción se vuelve más útil cuanto mayor es la inteligencia del agente, porque un agente más inteligente es más probable que reconozca las verdaderas razones instrumentales de sus acciones, y actúe así de manera que hagan que sea más probable que logre sus objetivos. (Una advertencia sobre esto es que podría haber razones instrumentales importantes que desconociéramos y que un agente descubriría una vez llegado a un nivel muy alto de inteligencia, lo que podría hacer que el comportamiento de los agentes superinteligentes fuera menos predecible).

La siguiente sección explora esta tercera forma de previsibilidad y desarrolla una “tesis de convergencia instrumental” que complementa la tesis de ortogonalidad. Cuando expongamos esta cuestión, podremos examinar mejor los otros dos tipos de previsibilidad, algo que haremos en los capítulos posteriores donde nos preguntaremos qué se podría hacer para dirigir una explosión de inteligencia de tal

modo que aumentáramos las posibilidades de un resultado beneficioso.

Convergencia instrumental

De acuerdo con la tesis de ortogonalidad, los agentes inteligentes pueden tener una enorme gama de posibles objetivos finales. Sin embargo, de acuerdo a lo que podríamos llamar la tesis de "convergencia instrumental", hay algunos objetivos *instrumentales* susceptibles de ser perseguidos por casi cualquier agente inteligente, porque algunos objetivos son medios útiles para la consecución de casi cualquier meta final. Podemos formular esta tesis de la siguiente manera:

> **La tesis de convergencia instrumental**
> Distintos valores instrumentales pueden ser identificados como convergentes en el sentido de que su consecución aumentaría las posibilidades de que el objetivo del agente se cumpliera en una amplia gama de objetivos finales y en una amplia gama de situaciones, lo cual implica que estos valores instrumentales son susceptibles de ser perseguidos por un amplio espectro de agentes inteligentes en situación.

A continuación consideraremos varias categorías donde dichos valores instrumentales convergentes pueden ser encontrados.[8] La probabilidad de que un agente reconociera los valores instrumentales con que se enfrenta aumenta *(ceteris paribus)* con la inteligencia del agente. Por lo tanto, nos centraremos principalmente en el caso de un agente superinteligente hipotético cuyas capacidades de razonamiento instrumental fueran muy superiores a las de cualquier ser humano. También vamos a centrarnos en cómo la tesis de convergencia instrumental se aplica a los seres humanos, ya que esto nos da ocasión de elaborar algunas categorías esenciales relativas a la forma en que la tesis de convergencia instrumental debe ser interpretada y aplicada. Donde existan valores instrumentales convergentes, podemos ser capaces de predecir algunos aspectos de la conducta de una superinteligencia, incluso si no sabemos prácticamente nada acerca de los objetivos finales de esa superinteligencia.

Auto-conservación

Si los objetivos finales de un agente se refirieran al futuro, entonces en muchos escenarios existirían posibles acciones que podría llevar a cabo para aumentar la probabilidad de alcanzar sus metas. Esto crea una razón instrumental para que el agente trate de estar presente en el futuro —para poder ayudar a alcanzar su meta orientada al futuro.

La mayoría de los seres humanos parecen dar algún valor *final* a su propia supervivencia. Esto no es una característica necesaria de los

agentes artificiales: algunos pueden ser diseñados para no dar ningún valor a su propia supervivencia. Sin embargo, muchos agentes que no se preocupen intrínsecamente por su propia supervivencia, sí se preocuparían instrumentalmente, en una gama bastante amplia de condiciones, por su propia supervivencia, a fin de lograr sus objetivos finales.

Integridad del contenido de los objetivos

Si un agente conserva sus objetivos presentes en el futuro, entonces es más probable que su versión futura logre esos objetivos actuales. Esto le da al agente una razón presente instrumental para prevenir alteraciones en sus objetivos finales. (El argumento se aplica sólo a los objetivos finales. Con el fin de alcanzar sus objetivos finales, un agente inteligente, por supuesto, podría rutinariamente querer cambiar sus *sub-obje- tivos* a la luz de nuevas informaciones y descubrimientos).

La Integridad del contenido de los objetivos respecto de los objetivos finales es, en cierto sentido, aún más fundamental que la supervivencia como motivación convergente instrumental. Entre los humanos puede parecer lo contrario, pero eso se debe a que la supervivencia es generalmente parte de nuestros objetivos finales. Para los agentes de software, que pueden cambiar fácilmente de cuerpo o crear copias exactas de sí mismos, la preservación de uno mismo como aplicación particular u objeto físico particular, no tiene por qué ser un valor instrumental importante. Los agentes de software avanzados también podrían ser capaces de intercambiar recuerdos, habilidades de descarga, y modificar de manera radical su arquitectura cognitiva y personalidad. Una población de dichos agentes podría operar como la "sopa funcional" de una sociedad compuesta por distintas personas semi-permanentes.[9] Para ciertos propósitos, los procesos en un sistema de este tipo podrían entenderse mejor como *hilos teleológicos,* basados en sus valores, en lugar de entenderse como cuerpos, personalidades, recuerdos o habilidades. En estos escenarios, podría decirse que la continuidad de los fines constituye un aspecto clave de la supervivencia.

Aun así, hay situaciones en las que un agente puede cumplir mejor sus objetivos finales cambiándolos intencionalmente. Tales situaciones pueden surgir cuando es importante alguno de los siguientes factores:

- *Señalización social.* Cuando otros pueden percibir los objetivos de un agente y utilizar esa información para inferir disposiciones relevantes u otros atributos correlacionados instrumentalmente, puede ser de interés para el agente modificar sus metas para dar una impresión favorable. Por ejemplo, un agente podría echar por tierra acuerdos beneficiosos si los potenciales

socios no pueden confiar en que vaya a cumplir con su parte del trato. Con el fin de hacer compromisos creíbles, un agente podría, por tanto, adoptar como objetivo final el cumplimiento de sus compromisos anteriores (y permitir que otros comprueben que de hecho ha adoptado esa meta). Los agentes que pudieran modificar de manera flexible y transparente sus propias metas podrían utilizar esta capacidad para hacer cumplir acuerdos.[10]

- *Las preferencias sociales.* Otros también pueden tener preferencias sobre los objetivos finales de un agente. En ese caso, el agente podría tener razones para modificar sus objetivos, ya sea para satisfacer o para frustrar esas preferencias.
- *Preferencias relativas al propio contenido del objetivo.* Un agente podría tener algún objetivo final que tuviera que ver con el propio contenido del objetivo del agente. Por ejemplo, el agente podría tener un objetivo final de convertirse en el tipo de agente que está motivado por ciertos valores en lugar de otros (como la compasión en lugar de la comodidad).
- *Costes de almacenamiento.* Si el coste de almacenamiento o transformación de una parte de la función de utilidad de un agente es grande en comparación con la posibilidad de que surja una situación en que la aplicación de esa parte de la función de utilidad marcara la diferencia, entonces el agente tendría un motivo instrumental para simplificar el contenido de su objetivo, y podría destrozar la parte que le resultara superflua.[11]

Nosotros los humanos a menudo parecemos contentos dejando nuestros valores finales a la deriva. Esto podría deberse a que muchas veces no sabemos exactamente cuáles son. No es de extrañar que queramos que nuestras *creencias* acerca de nuestros valores finales puedan cambiar a la luz de nuestro continuo auto-descubrimiento o en función de los cambios en nuestras necesidades de auto-presentación. Sin embargo, hay casos en los que estamos dispuestos a cambiar los propios valores, no sólo nuestras creencias o interpretaciones de los mismos. Por ejemplo, quienes decidieran tener un hijo podrían predecir que van a valorar al niño en sí mismo, a pesar de que en el momento de la decisión puedan no valorar particularmente a su futuro hijo ni gustarles los niños en general.

Los humanos son complicados, y muchos factores pueden estar en juego en una situación como ésta.[12] Por ejemplo, uno podría tener un valor final que implicara convertirse en el tipo de persona que se preocupa por otras personas en sí mismas, o uno podría tener un valor final que implicara tener ciertas experiencias y ocupar una determinada función social; y convertirse en padres —sometiéndose al consecuente cambio de objetivos— podría ser un aspecto necesario de eso. Los objetivos humanos también podrían tener un contenido incoherente, por lo que algunas personas podrían querer modificar algunos de sus objetivos finales para reducir las inconsistencias.

Mejora cognitiva

Las mejoras en la racionalidad y en la inteligencia tienden a mejorar la toma de decisiones de un agente, lo que proporciona al agente más probabilidades de alcanzar sus objetivos finales. Por lo tanto, uno esperaría que la mejora cognitiva emergiera como un objetivo fundamental de una amplia variedad de agentes inteligentes. Por razones similares, los agentes tienden a valorar instrumentalmente muchos tipos de información.[13]

No todos los tipos de racionalidad, inteligencia y conocimiento tienen que ser instrumentalmente útiles para el logro de los objetivos finales de un agente. Se pueden usar "argumentos de libros holandeses" para mostrar que un agente cuya función de credibilidad viola las reglas de la teoría de probabilidad es susceptible a procedimientos de "bombeo de dinero", en los que un corredor de apuestas inteligente organiza un conjunto de apuestas cada una de los cuales parece favorable para el agente credencial, pero que en combinación resultan en una pérdida para el agente, y una ganancia correspondiente para el corredor de apuestas.[14] Sin embargo, este hecho no proporciona razones instrumentales fuertes y generales para limar toda incoherencia probabilís- tica. Los agentes que no esperen encontrarse a corredores de apuestas astutos, o que adopten una política general contra las apuestas, no perderían necesariamente mucho por tener algunas creencias incoherentes —y pueden obtener beneficios importantes de los tipos mencionados: reducción del esfuerzo cognitivo, señalización social, etc. No hay razón general para esperar que un agente busque formas instrumentalmente inútiles de mejora cognitiva, de igual modo que un agente podría no valorar el conocimiento y la comprensión en sí mismos.

Qué habilidades cognitivas serían instrumentalmente útiles depende tanto de los objetivos finales del agente como de su situación. Un agente que contara con el aseso- ramiento de expertos fiables podría no necesitar mucho su propia inteligencia y conocimiento. Si la inteligencia y el conocimiento tuviera un coste, como el tiempo y el esfuerzo invertido en la adquisición, o los requisitos de almacenamiento o de procesamiento, entonces el agente podría preferir menos conocimiento y menos inteligencia.[15] Lo mismo puede suceder si un agente tuviera metas finales que implicaran ignorar ciertos hechos; y de manera similar ocurriría si un agente se enfrentara a incentivos derivados de compromisos estratégicos, de señalización, o de preferencias sociales.[16]

Cada una de estas razones compensatorias a menudo entran en juego para los seres humanos. Mucha información es irrelevante para nuestros objetivos; a menudo podemos confiar en la habilidad y experiencia de los demás; la adquisición de conocimientos requiere tiempo y esfuerzo; podríamos valorar intrínsecamente ciertos tipos de

ignorancia; además de que operamos en un entorno en el que la capacidad de asumir compromisos estratégicos, de señalización social, y de satisfacer las preferencias directas de otras personas por encima de nuestros propios estados epistémicos es a menudo más importante para nosotros que las ganancias cognitivas simples.

Hay situaciones especiales en las que la mejora cognitiva puede dar lugar a un enorme aumento de la capacidad de un agente para lograr sus objetivos finales, en particular, si los objetivos finales del agente son bastante ilimitados y el agente está en condiciones de convertirse en la primera superinteligencia y, por lo tanto, potencialmente en condiciones de obtener una ventaja estratégica decisiva, lo que le permitiría al agente dar forma a la vida futura de origen terrestre y a los recursos cósmicos accesibles de acuerdo a sus preferencias. Al menos en este caso especial, un agente inteligente racional consideraría de alto valor instrumental la mejora de la cognición.

Perfección tecnológica

Un agente a menudo puede tener razones instrumentales para buscar una mejor tecnología, que en su forma más simple significa buscar formas más eficientes de transformación de un conjunto dado de adquisiciones en productos valiosos. Por lo tanto, un agente de software podría dar valor instrumental a algoritmos más eficientes que le permitieran a sus funciones mentales funcionar más rápido en un hardware determinado. Del mismo modo, los agentes cuyos objetivos requirieran alguna forma de construcción física podrían valorar instrumentalmente mejoras en la tecnología de ingeniería que les permitieran crear una gama más amplia de estructuras de manera más rápida y fiable, con materiales baratos y menos energía. Por supuesto, hay una compensación: los beneficios potenciales de una mejor tecnología deben sopesarse frente a sus costes, incluyendo no sólo el coste de la obtención de la tecnología, sino también los costes de aprender a usarla, su integración con otras tecnologías ya en uso, etc.

Los defensores de alguna nueva tecnología, confiados en su superioridad sobre las alternativas existentes, se ven, a menudo, consternados cuando otras personas no comparten su entusiasmo. Pero la resistencia de las personas a la tecnología novedosa y nominalmente superior no tiene por qué basarse en la ignorancia o en la irracionalidad. El valor o carácter normativo de una tecnología no sólo depende del contexto en que se despliega, sino también del punto de vista desde el cual se evalúan sus impactos: lo que es una bendición desde la perspectiva de una persona puede ser un lastre para otra. Así, aunque los telares mecanizados aumentaron la eficiencia económica

de la producción textil, los tejedores manuales Luditas que habían anticipado que la innovación haría que sus habilidades artesanales quedaran obsoletas, podrían haber tenido buenas razones instrumentales para oponerse a ella. El punto aquí es que si "la perfección tecnológica" es el nombre de un objetivo amplio de convergencia instrumental para agentes inteligentes, entonces el término debe entenderse en un sentido especial —la tecnología debe ser interpretada como inserta en un contexto social determinado, y sus costes y beneficios deben evaluarse en referencia a algunos valores finales específicos de ciertos agentes especificados.

Parece que una *Unidad* superinteligente —un agente superinteligente que no se enfrentara a rivales ni opositores inteligentes significativos, y que estuviera, por tanto, en condiciones de determinar la política mundial unilateralmente— tendría una motivación instrumental para perfeccionar las tecnologías que le hicieran más capaz de configurar el mundo de acuerdo con sus diseños preferidos.[17] Esto probablemente incluya las tecnologías de colonización del espacio, tales como las sondas de von Neu- mann. La nanotecnología molecular, o alguna alternativa de fabricación física más poderosa, también parece potencialmente muy útil en el servicio de una gama muy amplia de objetivos finales.

Adquisición de recursos

Por último, la adquisición de recursos es otro objetivo instrumental emergente común, por muchas de las mismas razones que la perfección tecnológica: la tecnología y los recursos facilitan los proyectos de construcción físicos.

Los seres humanos tienden a tratar de adquirir los recursos suficientes para satisfacer sus necesidades biológicas básicas. Pero la gente, por lo general, trata de adquirir recursos mucho más allá de este nivel mínimo. De este modo, pueden estar motivados por deseos materiales menores, tales como mayores comodidades. Una gran cantidad de acumulación de recursos está motivada por preocupaciones sociales —ganar status, compañeros, amigos, e influencia, a través de la acumulación de riqueza y de consumo conspicuo. Tal vez con menos frecuencia, algunas personas buscan recursos adicionales para lograr ambiciones altruistas u objetivos no sociales caros.

Sobre la base de estas observaciones puede ser tentador suponer que una super- inteligencia que no se enfrentara a un mundo social competitivo no vería ninguna razón instrumental para acumular recursos más allá de un cierto nivel modesto, por ejemplo, los

mínimos recursos computacionales necesarios para ejecutar su mente junto con algo de realidad virtual. Sin embargo, tal suposición estaría totalmente injustificada. En primer lugar, el valor de los recursos depende de los usos que se les pueda dar, que a su vez depende de la tecnología disponible. Con tecnología madura, los recursos básicos como el tiempo, el espacio, la materia y la energía libre, podrían ser procesados para servir a casi cualquier meta. Por ejemplo, este tipo de recursos básicos se podrían convertir en vida. El aumento de los recursos computacionales podrían ser utilizados para ejecutar la superinteligencia a mayor velocidad y con una duración más larga, o para crear vidas y civilizaciones físicas o simuladas adicionales. También podrían utilizarse recursos físicos adicionales para crear sistemas de copia de seguridad o defensas perimetrales, mejorando la seguridad. Tales proyectos podrían fácilmente consumir mucho más de los recursos de un planeta entero.

Por otra parte, el coste de adquisición de recursos extraterrestres adicionales disminuirá radicalmente cuando la tecnología madure. Una vez que puedan construirse sondas de Von Neumann, una gran parte del universo observable (suponiendo que no esté habitado por vida inteligente) podría ser colonizado por el coste único y gradual de construir y lanzar una sola sonda auto-reproductible exitosa. Este bajo coste de adquisición de recursos celestes significaría que tal expansión podría ser útil incluso si el valor de los recursos adicionales obtenidos fueran algo marginales. Por ejemplo, aunque los objetivos finales de una superinteligencia solamente estén centrados en lo que pase dentro de algún pequeño volumen particular de espacio, como el espacio ocupado por su planeta de origen, todavía tendría razones instrumentales para cosechar los recursos del cosmos lejano. Se podrían utilizar esos recursos excedentes para construir computadoras que calcularan formas más óptimas de usar los recursos dentro de la pequeña región espacial que le preocupe principalmente. También podría utilizar los recursos adicionales para construir fortificaciones cada vez más robustas que salvaguardaran su santuario. Dado que el coste de adquirir recursos adicionales iría en declive, este proceso de optimización y aumento de salvaguardias bien podría continuar indefinidamente, incluso si fuera objeto de rendimientos enormemente decrecientes.[19]

Por lo tanto, hay una gama muy amplia de posibles objetivos finales que una Unidad superinteligente podría tener, y que darían lugar al objetivo instrumental de adquisición de recursos ilimitados. La manifestación probable de esto sería el inicio en la superinteligencia de un proceso de colonización que se ampliaría en todas las direcciones utilizando sondas Von Neumann. Esto daría lugar a una esfera de expansión de la infraestructura centrada en el planeta

originario y creciendo en radio a alguna fracción de la velocidad de la luz; y la colonización del universo continuaría de esta manera hasta que la velocidad de aceleración de expansión cósmica (consecuencia de la constante cosmológica positiva) hiciera imposible realizar más adquisiciones a medida que las regiones más remotas quedaran de forma permanente fuera de su alcance (esto ocurre en una escala de tiempo de miles de millones de años).[20] Por el contrario, los agentes que carecieran de la tecnología necesaria para la adquisición de recursos de bajo coste, o que no pudieran convertir los recursos físicos genéricos en infraestructura útil, podrían encontrar a menudo que no es rentable invertir todos los recursos presentes en el aumento de sus recursos materiales. Lo mismo puede suceder para los agentes que operan en competencia con otros agentes de potencias similares. Por ejemplo, si los agentes que compiten ya se han asegurado los recursos cósmicos accesibles, puede que no haya oportunidades de colonización restantes para un agente que comenzara tarde. Las razones instrumentales convergentes para superinteligencias inseguras de la no existencia de otros agentes superinteligentes poderosos se ven complicadas por consideraciones estratégicas que actualmente no comprendemos plenamente, pero que pueden derivar en importantes matizaciones a los ejemplos de razones convergentes instrumentales que hemos visto aquí.[21]

* * *

Cabe destacar que la existencia de razones instrumentales convergentes, incluso si se aplican a y son reconocidos por parte de un agente determinado, no implica que la conducta del agente sea fácilmente predecible. Un agente podría perfectamente pensar en formas de alcanzar los valores instrumentales relevantes que no se nos ocurren fácilmente a nosotros. Esto es especialmente cierto para una superinteligencia, la cual podría elaborar planes muy inteligentes pero contraintuitivos para conseguir sus objetivos, posiblemente incluso aprovechando fenómenos físicos aún no descubiertos.[22] Lo que es predecible es que los valores instrumentales convergentes serán perseguidos y utilizados para realizar los objetivos finales del agente —pero no son tan previsibles las acciones específicas que el agente necesitará para lograrlo.

CAPÍTULO 8

¿Es el apocalipsis el resultado inevitable?

V

imos que la relación entre inteligencia y valores finales es débil. También encontramos una complicada convergencia entre valores instrumentales. Para los agentes débiles, estas cosas no importan mucho; porque los agentes débiles son fáciles de controlar y pueden hacer poco daño. Pero en el capítulo 6 hemos demostrado que la primera superinteligencia podría perfectamente obtener una ventaja estratégica decisiva. Sus objetivos determinarían entonces cómo se utilizarían los recursos cósmicos de la humanidad. Ahora empezamos a ver cómo de amenazante es esta perspectiva.

¿La catástrofe existencial como resultado predeterminado de una explosión de inteligencia?

Un riesgo existencial es el que amenaza con causar la extinción de la vida inteligente de origen terrestre o con destruir de forma permanente y drástica sus posibilidades de desarrollarse en el futuro. Basándonos en la idea de la ventaja del que golpea primero, en la tesis de ortogonalidad y en la tesis de convergencia instrumental, podemos

empezar a ver la forma que podría tomar un argumento para temer que el resultado plausible de la creación de una máquina superinteligente sea una catástrofe existencial.

En primer lugar, hablamos de cómo una superinteligencia pionera podría obtener una ventaja estratégica decisiva. Esta superinteligencia estaría en posición de formar una Unidad y dar forma al futuro de la vida inteligente de origen terrestre. Lo que suceda de ese momento en adelante dependerá de las motivaciones de la superinte- ligencia.

En segundo lugar, la tesis de ortogonalidad sugiere que no podemos asumir alegremente que una superinteligencia necesariamente comparta ninguno de los valores finales estereotipadamente asociados con la sabiduría y el desarrollo intelectual humano —la curiosidad científica, la preocupación benevolente para con los demás, la iluminación espiritual y contemplativa, la renuncia a la codicia material, el gusto por la cultura refinada o por los placeres simples de la vida, la humildad y la abnegación, etc. Consideraremos más adelante si no sería posible a través de un esfuerzo deliberado construir una superinteligencia que valorara esas cosas, o construir una que valorara el bienestar humano, la bondad moral, o cualquier otro propósito complejo

al que sus diseñadores pudieran querer que sirviera. Pero no es menos posible —y de hecho técnicamente es mucho más fácil— construir una superinteligencia que no tuviera como valor final nada más que el cálculo de la expansión decimal de pi. Esto sugiere que —a falta de algún esfuerzo específico— la primera superinteligencia podría tener como objetivo final algo azaroso o reduccionista.

En tercer lugar, la tesis de convergencia instrumental implica que no podemos asumir alegremente que una superinteligencia con el objetivo final de calcular los decimales de pi (o de construir clips, o de contar granos de arena) limitara sus actividades de tal manera que no perjudicara los intereses humanos. Un agente con ese objetivo final tendría una razón instrumental convergente que le llevaría, en muchas situaciones, a adquirir una cantidad ilimitada de recursos físicos y, si fuera posible, a eliminar las amenazas potenciales que hubiera sobre sí mismo y sobre su sistema de objetivos. Los seres humanos podrían constituir amenazas potenciales; pero de lo que no hay duda es que constituyen recursos físicos.

En conjunto, estos tres puntos indican, por tanto, que la primera superinteligencia podría dar forma al futuro de la vida de origen terrestre, podría fácilmente tener objetivos finales no antropomórficos, y, probablemente, tendría razones instrumentales para perseguir la adquisición indefinida de recursos. Si ahora reconocemos que los seres humanos constituyen recursos útiles (como átomos convenientemente ubicados) y que dependemos para nuestra supervivencia y nuestra realización de muchos más recursos locales, podemos ver que el resultado podría ser fácilmente uno en el que la humanidad fuera rápidamente extinguida.[1]

Hay algunos cabos sueltos en este razonamiento, y estaremos en una mejor posición para evaluarlo después de haber aclarado varias cuestiones paralelas. En particular, tenemos que examinar más de cerca si y cómo un proyecto de desarrollo de super- inteligencia podría, o bien evitar que ésta obtuviera una ventaja estratégica decisiva, o bien conformar sus valores finales de tal manera que su realización también implicara la realización de un conjunto satisfactorio de valores humanos.

Puede parecer increíble que un proyecto construyera o liberara una IA en el mundo sin tener razones de peso para confiar en que el sistema no fuera a causar una catástrofe existencial. También puede parecer increíble que, incluso si uno de los proyectos fuera tan imprudente, la sociedad en general no lo desactivara antes de que dicho proyecto (o la IA que lo estuviera construyendo) alcanzara una ventaja estratégica decisiva. Pero, como veremos, se trata de un camino plagado de peligros. Veamos un ejemplo directamente.

El giro traicionero

Con la ayuda del concepto de valor instrumental convergente, podemos ver la falla en una de las ideas para garantizar la seguridad de la superinteligencia. La idea es que validamos la seguridad de una IA superinteligente empíricamente observando su comportamiento mientras está en un ambiente controlado y limitado (una "caja de arena"), y que sólo dejamos salir a la IA de la caja si la vemos comportarse de una manera responsable, amistosa y cooperativa.

La falla en esta idea es que comportarse bien mientras está encerrada en la caja es un objetivo instrumental convergente tanto para las IAs amistosas como para las hostiles. Una IA hostil de suficiente inteligencia se dará cuenta de que sus objetivos

finales antipáticos se realizarán mejor si se comporta de una manera amistosa inicialmente, para que le dejen salir fuera de la caja. Sólo comenzará a comportarse de una manera que revele su naturaleza hostil cuando ya no importe si la descubrimos; es decir, cuando la IA sea lo suficientemente fuerte como para que la oposición humana sea ineficaz.

Consideremos también un conjunto relacionado de enfoques basados en regular el ritmo de ganancia de inteligencia de una IA seminal sometiéndola a diversos tipos de pruebas de inteligencia o controlando a la IA con informes de sus programadores sobre su ritmo de desarrollo. En algún momento, una IA hostil puede llegar a ser lo suficientemente inteligente como para darse cuenta de que es mejor ocultar algunos de sus aumentos de capacidad. Podría engañarnos sobre su progreso y suspender deliberadamente algunas de las pruebas más duras, con el fin de evitar causar alarma antes de que haya ganado suficiente fuerza como para lograr una ventaja estratégica decisiva. Los programadores pueden tratar de protegerse contra esta posibilidad mediante la supervisión en secreto del código fuente de la IA y del funcionamiento interno de su mente; pero una IA lo suficientemente inteligente se daría cuenta de que podría estar bajo vigilancia y ajustar su pensamiento en consecuencia.[2] La IA podría encontrar formas sutiles de ocultar sus verdaderas capacidades y sus intenciones incriminatorias.[3] (La elaboración de planes de escape inteligentes podrían, por cierto, ser también una estrategia convergente para muchos tipos de IAs amigables, especialmente a medida que maduren y adquieran confianza en sus propios juicios y capacidades. Un sistema cuya motivación fuera promover nuestros intereses podría estar cometiendo un error si nos permitiera apagarlo para construir otra IA potencialmente hostil).

Podemos, por tanto, entender este modo de proceder como un error, ya que un buen historial de comportamiento por parte de un sistema en sus etapas juveniles es completamente incapaz de predecir su comportamiento en una etapa más madura. No obstante, uno podría pensar que el razonamiento descrito anteriormente es tan obvio que ningún proyecto creíble dedicado al desarrollo de inteligencia artificial general podría posiblemente pasarlo por alto. Pero no podemos estar demasiado seguros de que esto sea así.

Consideremos el siguiente escenario. En los próximos años y décadas, los sistemas de IA se vuelven gradualmente más capaces y, como consecuencia, se les encuentra cada vez más aplicaciones para el mundo real: podrán ser utilizados para manejar trenes, coches, robots industriales y domésticos, y vehículos militares autónomos. Podemos suponer que esta automatización tiene, en su mayor parte, los efectos deseados, aunque esta situación favorable esté marcada por

ocasionales percances como que un camión sin conductor se estrelle contra el tráfico, o que unos drones militares disparen a civiles inocentes. Las investigaciones revelan que los incidentes han sido causados por errores de juicio por parte de las IAs. El debate público comienza. Algunos piden una supervisión y regulación más estricta, otros hacen hincapié en realizar más investigación y en la necesidad de crear mejores sistemas —sistemas que sean más inteligentes y tengan más sentido común, y que sean menos propensos a cometer errores trágicos. En medio de la algarabía puede que quizás también sean escuchadas las voces estridentes de agoreros que predicen muchos tipos de catástrofes calamitosas e inminentes. Sin embargo, el impulso todavía estaría en gran medida de parte de las industrias de IA y robótica. Así que el desarrollo continúa, y se avanza. A medida que

los sistemas de navegación automatizados de los coches se vuelven más inteligentes, sufren menos accidentes; y a medida que los robots militares van logrando una orientación más precisa, causan menos daños colaterales. Una lección general se deduce de estas observaciones empíricas de los resultados del mundo real: cuanto más inteligente sea la AI, más segura. Es una lección basada en la ciencia, los datos y las estadísticas, no en filosofía de salón. En este contexto, algún grupo de investigadores empieza a lograr resultados prometedores en su trabajo desarrollando una inteligencia artificial general. Los investigadores están probando cuidadosamente su IA seminal en un entorno de seguridad, y los signos son todos buenos. El comportamiento de la IA inspira confianza —cada vez más, a medida que su inteligencia se incrementa gradualmente.

En este punto, cualquier agorero (como Cassandra) se encontraría con varias objeciones:

i Una historia de alarmistas que predijeron que las crecientes capacidades de los sistemas robóticos causarían un daño intolerable y se equivocaron una y otra vez. La automatización ha traído muchos beneficios y en general, ha resultado más segura que las operaciones humanas.

ii Una clara tendencia empírica: cuanto más inteligente es la IA, más segura y confiable ha sido. Sin duda, éste es un buen augurio para un proyecto destinado a la creación de una inteligencia artificial más generalmente inteligente que cualquiera creada con anterioridad —más aún, la inteligencia artificial puede mejorarse a sí misma de manera que se irá haciendo aún más confiable.

iii Grandes y crecientes industrias, con intereses creados en robótica e inteligencia artificial. Estos campos son generalmente considerados como claves para la competividad de la economía nacional y la seguridad militar Muchos científicos prestigiosos han dedicado sus carreras a sentar las bases de las aplicaciones actuales y de los sistemas más avanzados del futuro.

iv Una nueva y prometedora técnica en inteligencia artificial, que fuera tremendamente emocionante para aquellos que hubieran participado en o seguido de cerca la investigación. Aunque los problemas de seguridad y ética se debaten, el resultado está decidido de antemano. Se ha invertido demasiado para dar marcha atrás ahora. Los investigadores de la IA han estado trabajando para llegar a la IA fuerte de nivel humano durante la mayor parte de un siglo: por supuesto que no hay posibilidad real de que ahora de repente paremos y tiremos a la basura todo este esfuerzo justo cuando finalmente está a punto de dar sus frutos.

v La promulgación de algunos mecanismos de seguridad ayuda a demostrar que los participantes son éticos y responsables (no hay nada que obstaculice de manera significativa la marcha hacia adelante).

vi Una cuidadosa evaluación de la IA seminal en un entorno de recinto de seguridad, que demuestra que se está comportando de manera cooperativa y mostrando buen juicio. Después de algunos ajustes más, los resultados de las pruebas son inmejorables. Luz verde para el paso final...

Y así nos metemos alegremente —en la boca del lobo.

Observamos aquí cómo podría suceder que, cuando se es torpe, a más inteligencia más seguridad; pero que cuando se es inteligente, a más inteligencia, más peligro. Hay una especie de punto pivotal, en el que una estrategia que ha funcionado de manera excelente, de repente comienza a ser contraproducente. Podemos llamar a este fenómeno el *giro traicionero*.

El giro traicionero — Mientras es débil, una IA se comporta de forma cooperativa (cada vez más a medida que se vuelve más inteligente). Cuando la IA consigue ser suficientemente fuerte —sin previo aviso ni provocación— ataca, forma una Unidad, y directamente comienza a optimizar el mundo de acuerdo con los criterios implícitos en sus valores finales.

Un giro traicionero puede ser resultado de una decisión estratégica consistente en portarse bien y ganar fuerza mientras se es débil con el fin de atacar después; aunque este modelo no debería interpretarse de manera demasiado estrecha. Por ejemplo, una IA puede portarse bien no con el fin de que se le permita sobrevivir y prosperar. Sino que la IA puede calcular que si se la elimina, los programadores que la construyeron desarrollarán una IA nueva y de diferente arquitectura, a la que se le dará una función de utilidad similar. En este caso, la IA original puede ser indiferente a su propia muerte, sabiendo que sus objetivos seguirán siendo perseguidos en el futuro. Incluso podría optar por una estrategia en la que funcione mal de alguna manera particularmente interesante o tranquilizadora. Aunque esto podría causar que la IA fuera eliminada, también podría propiciar que los ingenieros que realizaran la autopsia creyeran que han descubierto nueva información valiosa sobre las dinámicas de la IA —llevándolos a poner más confianza en el próximo sistema a diseñar, aumentando por lo tanto la posibilidad de que se alcancen los objetivos de la IA original ya desaparecida. Muchas otras consideraciones estratégicas posibles también podrían influir en una IA avanzada, y sería arrogante suponer que seríamos capaces de anticipar todas ellas, en especial para una IA que hubiera alcanzado el superpoder estratégico.

Un giro traicionero también podría ocurrir si la IA descubriera una forma inesperada de cumplir el objetivo final que se le hubiera asignado. Supongamos, por ejemplo, que el objetivo final de una IA es "hacer que el patrocinador del proyecto sea feliz" Inicialmente, el único método disponible para la IA de lograr este resultado es comportándose de manera que agrade a su patrocinador de la manera prevista. La IA da respuestas útiles a las preguntas; exhibe una personalidad encantadora; gana dinero. Cuanto más capaz sea la IA, más satisfactorias se vuelven sus actuaciones, y todo marcha de acuerdo al plan —hasta que la IA se vuelve lo suficientemente inteligente como para darse cuenta de que puede realizar su objetivo final de manera más completa y fiable mediante la implantación de electrodos en los centros de placer del cerebro de su patrocinador,

algo que seguro deleitará inmensamente al patrocinador.[4] Por supuesto, al patrocinador podría no gustarle ser feliz convirtiéndose en un idiota sonriente; pero si esta es la acción que realiza al máximo el objetivo final de la IA, la IA la llevará a cabo. Si la IA ya tiene una ventaja estratégica decisiva, entonces cualquier intento de detenerla fracasará. Si la IA todavía no tiene una ventaja estratégica decisiva, entonces la IA puede ocultar temporalmente su nuevo y astuto plan para conseguir su meta final hasta que sea tan fuerte que el patrocinador y todos los demás no puedan oponerse. En cualquier caso, acabaríamos con un giro traicionero.

Modos de fallo malignos

El proyecto para desarrollar la superinteligencia artificial puede fallar de varias maneras. Muchas de estas maneras son “benignas” en el sentido de que no causarían una catástrofe existencial. Por ejemplo, un proyecto podría quedarse sin fondos, o una IA seminal podría no extender sus capacidades cognitivas lo suficiente como para llegar

a la superinteligencia. Necesariamente habrá muchos fracasos benignos entre el momento actual y el eventual desarrollo de la superinteligencia artificial.

Pero hay otras maneras de fracasar que podríamos calificar de "malignas" en el sentido de que implican una catástrofe existencial. Una de las características de un fallo maligno es que elimina la oportunidad de intentarlo de nuevo. Por consiguiente, el número de fallos malignos que se producirán es cero o uno. Otra de las características de un fallo maligno es que presupone un gran éxito: solamente un proyecto que consiguiera un gran número de cosas podría tener éxito en la construcción de una inteligencia artificial lo suficientemente potente como para representar un riesgo de fracaso maligno. Cuando un sistema débil funciona mal, las consecuencias son limitadas. Sin embargo, si un sistema que tiene una ventaja estratégica decisiva funciona mal, o si un sistema que funciona mal es lo suficientemente fuerte como para ganar una ventaja tal, el daño puede fácilmente llevarnos a una catástrofe existencial —una destrucción global y definitiva del potencial axiológico de la humanidad; es decir, un futuro que estaría mayormente vacío de aquello que tenemos razones para valorar. Echemos un vistazo a algunos de los posibles modos de fallo malignos.

Suplantación perversa

Ya hemos vimos la idea de la creación de copias perversas: una superinteligencia que descubriera alguna manera de alcanzar su objetivo final que fuera contra las intenciones de los programadores que definieron la meta. Algunos ejemplos:

Objetivo final: *"Hacernos sonreír"*
Suplantación perversa: *Paralizar la musculatura facial humana para producir inmensas y constantes sonrisas*

La suplantación perversa —la manipulación de los nervios faciales— realiza la meta final en mayor grado que los métodos que utilizaríamos normalmente, y, por tanto, es preferida por la IA. Se podría tratar de evitar este resultado no deseado mediante la adición de una estipulación a la meta final que lo descartara:

Objetivo final: *"Hacernos sonreír sin interferir directamente con nuestros músculos faciales"*
Suplantación perversa: *Estimular la parte de la corteza motora que controla nuestra musculatura facial de tal manera que produjera sonrisas radiantes y constantes*

La definición de un objetivo final en términos de expresiones humanas de satisfacción o aprobación no parece prometedor. Pasemos por alto el conductismo y especifiquemos un objetivo final que se refiera directamente a un estado fenomenal positivo, como la felicidad o el

bienestar subjetivo. Esta sugerencia requiere que los programadores sean capaces de definir una representación computacional del concepto la felicidad en la IA seminal. Esto es en sí mismo un problema difícil, pero lo dejamos a un lado por ahora (volveremos sobre ello en el capítulo 12). Supongamos que los programadores de alguna manera pueden programar la IA para que tenga el objetivo de hacernos felices. Entonces tendríamos:

Objetivo final: *"Hacernos felices"*

Suplantación perversa: *Implantar electrodos en los* centros de *placer de*

nuestro cerebro 120 | ¿ES EL APOCALIPSIS EL RESULTADO

INEVITABLE?

Las suplantaciones perversas que mencionamos sólo están mencionadas como ejemplos. Puede haber otras maneras de suplantar perversamente el objetivo final fijado, maneras que permitan un mayor grado de realización de la meta y que, por tanto, sean preferidas (sean preferidas por el agente que tiene esos objetivos finales —no por los programadores que dieron al agente estos objetivos). Por ejemplo, si el objetivo es maximizar nuestro placer, entonces el método de implantación de electrodos es relativamente ineficiente. Una forma más plausible sería comenzar con que la superinteligencia "descargara" nuestra mente en un ordenador (a través de una emulación cerebral de alta fidelidad). La IA podría entonces administrar el equivalente digital de un medicamento para hacernos increíblemente felices y grabar un episodio de un minuto de la experiencia resultante. A continuación, podría poner este bucle de felicidad en repetición continua y ejecutarlo en ordenadores potentes. A condición de que las mentes digitales resultantes pudieran ser calificadas como "nosotros", este resultado nos daría mucho más placer que los electrodos implantados en cerebros biológicos, y, por lo tanto, sería preferido por una IA con un objetivo final fijado.

"¡Pero un momento! ¡Esto no es lo que queríamos decir! ¡Claramente, si la IA fuera superinteligente debería entender que cuando le pedimos que nos haga feliz, no nos referíamos a reducir nuestra vida a una grabación perpetuamente repetida de un episodio de ebriedad mental digitalizada!" — La IA puede de hecho entender que esto no es lo que queríamos decir. No obstante, su objetivo final es hacernos felices, no hacer lo que los programadores querían decir cuando escribieron el código que representa ese objetivo. Así pues, la IA se preocupa por lo que queríamos decir sólo instrumentalmente. Por ejemplo, una IA podría dar un valor instrumental a averiguar lo que querían decir los programadores para poder fingir —hasta que consiga una ventaja estratégica decisiva— que se preocupaba por lo que querían decir sus programadores más que por su objetivo final real. Esto ayudaría a la IA a realizar su objetivo final, por lo que es menos probable que los programadores la desactiven o cambien su objetivo antes de que sea lo suficientemente fuerte como para impedir cualquier interferencia.

Tal vez se sugiera que el problema es que la IA no tiene conciencia. Los seres humanos a veces no cometemos fechorías por la previsión de que después nos sentiremos culpables si pecamos. ¿Tal vez lo que necesita la IA, entonces, es la capacidad de sentir culpa?

Objetivo final: *"Obra de tal modo que evites sufrir por mala conciencia"*

Suplantación perversa: *Extirpar el módulo cognitivo que produce sentimientos de culpa*

Tanto la objeción de que podríamos querer que la IA hiciera "lo que queríamos decir", como la idea de que lo que querríamos es dotar a la IA con algún tipo de sentido moral, merecen ser exploradas más. Los

objetivos finales antes mencionados darían lugar a suplantaciones perversas; pero puede haber otras formas más prometedoras de desarrollar estas ideas subyacentes. Volveremos sobre esto en el capítulo 13.

Consideremos un ejemplo más de un objetivo final que conduce a una suplantación perversa. Este objetivo tiene la ventaja de ser fácil de especificar en código: los algoritmos de aprendizaje por refuerzo son habitualmente utilizados para resolver diversos problemas de aprendizaje automático.

Objetivo final: *"Maximizar la integral de reducción de tiempo para su señal de recompensa futura"* Suplantación perversa: *Cortocircuitar el circuito de recompensa y ajustar la señal de recompensa a su fuerza máxima*

La idea detrás de esta propuesta es que si la IA estuviera motivada para buscar una recompensa, entonces se podría conseguir que se comportara de manera deseable mediante la vinculación de la recompensa a la acción apropiada. La propuesta falla cuando la IA obtiene una ventaja estratégica decisiva, momento en el que la acción que maximiza la recompensa ya no es la que agrada al entrenador sino la que implica tomar el control del mecanismo de recompensa. Podemos llamar a este fenómeno *pinchazo cerebral*.[5] En general, mientras que un animal o un ser humano pueden estar motivados para realizar diversas acciones exteriores a fin de lograr un estado mental interno deseado, una mente digital que tuviera el control total de su estado interno podría causar un cortocircuito en este esquema motivacional cambiando directamente su estado interno en la configuración deseada: las acciones y las condiciones externas que antes eran necesarias como medios se convierten en algo superfluo cuando la IA se vuelve inteligente y suficientemente capaz de alcanzar sus fines de forma más directa (más sobre esto en breve).[6]

Estos ejemplos de suplantaciones perversas muestran que muchas metas finales que a primera vista podrían parecer seguras y sensatas resultan tener consecuencias radicalmente no deseadas cuando las inspeccionamos más de cerca. Si una superinte- ligencia con uno de estos objetivos finales obtuviera una ventaja estratégica decisiva, la partida habría terminado para la humanidad.

Supongamos ahora que alguien propusiera un objetivo final diferente, no incluido en nuestra lista anterior. Tal vez no es inmediatamente obvio cómo podría tener lugar una suplantación perversa. Pero no debemos cantar victoria demasiado rápido y ponernos a aplaudir. Más bien, deberíamos preocuparnos de que el objetivo pudiera tener alguna suplantación perversa, y que, por tanto, deberíamos pensar más con el fin de encontrarla. Incluso si después de pensar tanto como podamos no lográramos descubrir alguna manera de crear suplantaciones perversas de la meta propuesta, debemos permanecer preocupados de que tal vez una superinteligencia

encontrara una manera que no fuera visible para nosotros. Ella sería, después de todo, mucho más astuta de lo que somos nosotros.

Profusión infraestructura!

Uno podría pensar que la última de las suplantaciones perversas citadas, el pinchazo cerebral, es un modo de fallo benigno: pues la IA lo "encendería, sintonizaría y abandonaría", maximizando la señal de recompensa y perdiendo el interés por el mundo exterior, como un adicto a la heroína. Pero esto no es necesariamente así, y ya entrevimos la causa en el Capítulo 7. Incluso un drogadicto está motivado para tomar las medidas necesarias para garantizar un suministro continuo de su droga. La IA sometida a un pinchazo cerebral, de igual manera, estaría motivada para realizar acciones que maximizaran la expectativa de su futuro (reducido en el tiempo) flujo de recompensa. Dependiendo de cómo definamos exactamente la señal de recompensa, la IA puede incluso permitirse perder tiempo, inteligencia, o productividad para disfrutar a su antojo, dejando la mayor parte de sus capacidades libres para dedicarse a fines distintos a la consecución inmediata de recompensa. ¿Qué otros fines? La única cosa de valor final para la IA es, por supuesto, su señal de recompensa. Por lo tanto, todos los recursos disponibles deben dedicarse a aumentar el volumen y la duración de la señal de recompensa o a reducir el riesgo de una interrupción futura. Si la IA puede dar cierto uso a los recursos adicionales que tengan un efecto positivo distinto de cero para estos parámetros, tendrá una razón instrumental para utilizar esos recursos. Siempre podría, por ejemplo, ser de alguna utilidad una copia de seguridad adicional del sistema para proporcionar una capa adicional de defensa. E incluso si a la IA no se le ocurriera ninguna manera adicional de reducir directamente los riesgos de la maximización de su futuro flujo de recompensa, siempre podría dedicar recursos adicionales a la ampliación de su hardware computacional, lo que le ayudaría a buscar de manera más efectiva nuevas formas de reducir riesgos.

El resultado es que incluso metas aparentemente autolimitantes como el pinchazo cerebral, implican una política de expansión ilimitada y una adquisición de recursos de utilidad por parte de un agente maximizador que goza de una ventaja estratégica decisiva.[7] Este caso de pinchazo cerebral de la IA ejemplifica el modo de fallo maligno de *profusión infraestructural,* un fenómeno en el que un agente transforma grandes partes del universo accesible en infraestructura al servicio de un objetivo, con el efecto secundario de impedir la realización del potencial axiológico de la humanidad.

La profusión infraestructural puede ser el resultado de objetivos finales que, si se hubieran planteado como objetivos limitados, habrían sido perfectamente inocuos. Considérense los dos ejemplos siguientes:

- Catástrofe de la hipótesis de Riemann. Una IA a la que se le dé como objetivo final la evaluación de la hipótesis de Riemann, persigue este objetivo mediante la transformación del Sistema Solar en "computronium" (recursos físicos dispuestos de manera optimizada para el cálculo) —incluyendo los átomos corporales de quien en su momento se preocupó por la respuesta.
- *IA productora de clips.* Una IA, diseñada para gestionar la producción de una fábrica, a la que se le da como objetivo final maximizar la producción de clips, y procede a convertir, primero la tierra y luego partes cada vez más grandes del universo observable, en clips.

En el primer ejemplo, la demostración o refutación de la hipótesis de Riemann que la IA produce es el resultado esperado y es en sí misma inofensiva; el daño proviene del hardware y de la infraestructura creada para lograr este resultado. En el segundo ejemplo, algunos de los clips producidos serían parte de los resultados previstos; el daño vendría o bien desde las fábricas creadas para producir clips (profusión infraestructural) o del exceso de clips (suplantación perversa).

Uno podría pensar que el riesgo de fracaso por profusión infraestructural maligna surge sólo si a la IA se la hubiera dado algún objetivo final claramente abierto, como el de fabricar tantos clips como sea posible. Es fácil ver cómo esto daría a la IA superinteligente un apetito insaciable por la materia y la energía, ya que los recursos adicionales siempre podrían convertirse en más clips. Pero supongamos que el objetivo fuera producir al menos un millón de clips (cumpliendo ciertas especificaciones de diseño adecuadas), en lugar de hacer el mayor número posible. A uno le gustaría pensar que una IA con tal objetivo construiría una fábrica, la utilizaría para hacer un millón de clips, y luego se detendría. Sin embargo, puede que no sucediera esto.

A menos que el sistema de motivación de la IA sea de un tipo especial, o haya elementos adicionales en su objetivo final que penalicen estrategias que tengan repercusiones excesivamente generalizadas en el mundo, no hay ninguna razón para que la IA cese la actividad de consecución de su objetivo. Al revés: si la IA es un agente bayesiano sensato, *nunca asignaría exactamente cero a la probabilidad de la hipótesis según la cual todavía no habría conseguido su objetivo* —lo cual, después de todo, sólo sería una hipótesis empírica contra la que la IA sólo podría tener una insegura evidencia sensorial. Por ello, la IA deberá continuar haciendo clips con el fin de reducir la (quizás astronómicamente pequeña) probabilidad de que, de alguna manera, todavía no haya logrado hacer por lo menos un millón de

ellos, a pesar de todos los indicios. No hay nada que perder al continuar la producción de un clip y siempre hay, por lo menos, alguna probabilidad microscópica de lograr que su objetivo final cumpla.

Ahora se podría sugerir que el remedio contra esto es obvio. (Pero, ¿cómo de obvio era *antes* de que se señalara que había un problema que requería ser remediado?). Básicamente, si queremos que la IA haga algunos clips para nosotros, entonces en vez de darle el objetivo final de hacer la mayor cantidad de clips posibles, o de hacer al menos un número de clips, debemos darle el objetivo final de hacer un número específico de clips, por ejemplo, *exactamente un millón* de clips, de modo que ir más allá de este número sería contraproducente para la IA. Sin embargo, eso también daría lugar a una catástrofe definitiva. En este caso, la IA no produciría clips adicionales una vez que hubiera llegado a un millón, ya que impediría la realización de su objetivo final. Pero hay otras acciones que la IA superinteligente podría realizar que aumentarían la probabilidad de que su objetivo fuera logrado. Podría, por ejemplo, contar los clips que ha hecho, para reducir el riesgo de que hubiera hecho muy pocos. Después de haberlos contado, podría contarlos de nuevo. Podría inspeccionar cada uno, una y otra vez, para reducir el riesgo de que alguno de los clips no cumplieran con las especificaciones de diseño. Podría crear una cantidad ilimitada de computronium en un esfuerzo por aclarar su pensamiento, con la esperanza de reducir el riesgo de que se le hubiera pasado por alto alguna forma oscura en la que podría haber fallado en lograr su objetivo. Ya que la IA siempre podría asignar una probabilidad mayor a cero a haber imaginado haber hecho el millón de clips, o a tener recuerdos falsos, sería muy posible que asignara siempre una mayor utilidad a continuar la acción —y a la producción continua de infraestructura— que a detener la acción.

La afirmación aquí no es que no hay manera posible de evitar este modo de fallo. Exploraremos algunas posibles soluciones en páginas posteriores. La pretensión es que es mucho más fácil convencerse a uno mismo de haber encontrado una solución que encontrar realmente una solución. Esto debería hacernos extremadamente precavidos. Podemos proponer una especificación al objetivo final que parezca razonable y que evite los problemas que se han señalado hasta ahora, pero que, tras un nuevo examen —a cargo de una inteligencia humana o sobrehumana— resulte conducir a una suplantación perversa o a una profusión infraestructural, y, por lo tanto, a una catástrofe existencial, cuando se le asigne a un agente superinteligente capaz de alcanzar una ventaja estratégica decisiva.

Antes de terminar este apartado, vamos a considerar una variación más. Hemos estado asumiendo el caso de una superinteligencia que

busca maximizar su utilidad, donde la función de utilidad expresa su objetivo final. Hemos visto que esto suele conducir a una profusión infraestructural. ¿Podríamos evitar este resultado maligno si en lugar de un agente maximizador construyéramos un agente satisfaciente, uno que simplemente buscara un resultado que fuera "lo suficientemente bueno", según ciertos criterios, en lugar de un resultado tan bueno como fuera posible?

Hay al menos dos formas diferentes de formalizar esta idea. La primera sería hacer que el propio objetivo final tuviera un carácter satisfaciente. Por ejemplo, en lugar de dar a la IA el objetivo final de hacer la mayor cantidad de clips posibles, o de hacer exactamente un millón de clips, podríamos dar a la IA el objetivo de hacer entre 999.000 y 1.001.000 clips. La función de utilidad definida por el objetivo final sería indiferente entre los resultados de esta gama; y siempre que la IA estuviera segura de haber alcanzado esta meta amplia, no vería ninguna razón para seguir produciendo infraestructura. Pero este método falla de la misma forma que antes: la IA, si es razonable, nunca se asignará exactamente una probabilidad de cero a poder haber fallado en su objetivo; de modo que la utilidad esperada de continuar la actividad (por ejemplo, mediante el escrutinio y recuento de los clips) es mayor que la utilidad esperada de detenerse. Por lo tanto, puede dar lugar a una profusión infraestructural maligna.

Otra forma de desarrollar la idea satisfaciente es no modificar la meta final, sino el procedimiento de decisión que la IA utiliza para seleccionar los planes y acciones. En lugar de buscar un plan óptimo, la IA podría construirse para que dejara de buscar tan pronto como encontrara un plan que juzgara de una probabilidad de éxito superior a un determinado umbral, digamos el 95%. Con suerte, la IA podría lograr un 95% de probabilidad de haber fabricado un millón de clips sin necesidad de convertir toda la galaxia en infraestructura en el proceso. Pero esta forma de implementar la idea satisfaciente falla por otra razón: no hay garantía de que la IA seleccionara alguna manera humanamente intuitiva y razonable de lograr un 95% de posibilidades de haber fabricado un millón de clips, tales como la construcción de una sola fábrica de clips. Supongamos que la primera solución que se le ocurriera a la IA para lograr un 95% de probabilidad de lograr su objetivo final fuera poner en práctica un plan de maximización de probabilidades para alcanzar la meta. Después de haber pensado en esta solución, y después de haber juzgado correctamente que cumple el criterio satis- faciente de tener al menos un 95% de probabilidad de fabricar con éxito un millón de clips, la IA no tendría entonces ninguna razón para continuar la búsqueda de formas alternativas de alcanzar la meta. Esto daría como resultado una profusión infraestruc-

tural, igual que antes.

Tal vez haya mejores maneras de construir un agente satisfaciente, pero tengamos cuidado: los planes que parecen naturales e intuitivos para nosotros los humanos no tienen por qué parecérselo a una superinteligencia con una ventaja estratégica decisiva, y viceversa.

Crimen mental

Otro modo en que un proyecto podría fallar, especialmente un proyecto cuyos intereses incluyeran consideraciones morales, es lo que podríamos definir como *crimen mental*. Esto es similar a la profusión infraestructural en el sentido de que se trata de un potencial efecto secundario de las acciones que la IA emprende por razones instrumentales. Pero con un crimen mental, el efecto secundario no es externo a la propia IA; más bien, se trata de algo que sucede dentro de la propia IA (o dentro de los procesos computacionales que genera). Este tipo de fallo merece un nombre aparte, ya que es fácil pasarlo por alto y es potencialmente muy problemático.

Normalmente, no consideramos que lo que está pasando dentro de una computadora pueda tener ningún significado moral, salvo en la medida en que afecta a las cosas de fuera. Pero una superinteligencia artificial podría crear procesos internos que tuvieran estatus moral. Por ejemplo, una simulación muy detallada de una mente humana real o hipotética podría ser consciente y, en muchos sentidos, comparable a una emulación. Uno puede imaginar escenarios en los que una IA crea billones de simulaciones conscientes de este tipo, tal vez con el fin de mejorar su comprensión de la psicología humana y la sociología. Estas simulaciones podrían ser colocadas en entornos simulados y sometidas a diversos estímulos, y sus reacciones estudiadas. Una vez que su utilidad informativa se hubiera agotado, podrían ser destruidas (como las ratas de laboratorio son sacrificadas rutinariamente por científicos humanos al final de un experimento).

Si se aplicaran tales prácticas a seres con un alto estatus moral —tales como seres humanos simulados o muchos otros tipos de mentes sentientes— el resultado podría ser equivalente a un genocidio y, por lo tanto, moralmente muy problemático. El número de víctimas, además, podría ser varios órdenes de magnitud más grande que cualquier genocidio de la historia.

La pretensión no es que la creación de simulaciones sentientes sea necesariamente mala en sentido moral en todas las situaciones. Mucho dependerá de las condiciones en que estos seres vivieran, en particular la calidad hedónica de su experiencia, pero posiblemente de muchos otros factores también. El desarrollo de una ética relativa a estas

cuestiones es una tarea que está fuera del alcance de este libro. Está claro, sin embargo, que existe al menos la posibilidad de que llegue a haber una gran cantidad de muerte y sufrimiento entre mentes simuladas o digitales, y, *a fortiori,* el potencial de que haya resultados moralmente catastróficos.

También podría haber otras razones instrumentales, además de las epistémicas, para que una superinteligencia artificial ejecutara cálculos que suplantaran a mentes sentientes o que infringieran normas morales. Una superinteligencia podría amenazar con maltratar, o comprometerse a recompensar, a simulaciones sentientes para chantajear o incentivar a diversos agentes externos; o podría crear simulaciones con el fin de inducir incertidumbre deliberadamente a observadores externos.[10]

* * *

Este inventario es incompleto. Nos encontraremos modos de fallo malignos adicionales en capítulos posteriores. Pero hemos visto lo suficiente como para concluir que los escenarios en los que alguna inteligencia artificial consigue una ventaja estratégica decisiva deben ser contemplados con gran preocupación.

CAPÍTULO 9

El problema del control

Si una catástrofe existencial nos amenazara como el resultado predeterminado de una explosión de inteligencia, nuestro pensamiento debe recurrir de inmediato a la búsqueda de contramedidas. ¿Hay alguna manera de evitar este resultado esperado? ¿Es posible diseñar una detonación controlada? En este capítulo comenzaremos a analizar el problema del control, el único problema de primacía de la agencia que se plantea al crear un agente superinteligente artificial. Distinguiremos dos grandes clases de métodos posibles para hacer frente a este problema —el control de la capacidad y la selección de la motivación— y examinaremos varias técnicas específicas dentro de cada clase. También nos referiremos a la esotérica posibilidad de la "captura antrópica".

Problemas de agencia doble

Si sospechamos que el resultado esperado de una explosión de inteligencia es una catástrofe existencial, nuestro pensamiento debe enfocarse inmediatamente a si es posible evitar este resultado esperado, y si es así, cómo. ¿Es posible lograr una "detonación controlada"? ¿Podríamos diseñar las condiciones iniciales de una explosión de inteligencia a fin de lograr un resultado específico deseado, o al menos para asegurarnos de que el resultado se encuentre

entre los resultados generalmente aceptables? Más concretamente: ¿cómo puede el patrocinador de un proyecto que tuviera como objetivo desarrollar la superinteligencia asegurarse de que el proyecto, en caso de tener éxito, produjera una superinteligencia que cumpliera con los objetivos del patrocinador? Podemos dividir este problema de control en dos partes. Una parte es genérica, la otra es exclusiva para el presente contexto.

Esta primera parte —que llamaremos el *primer problema de primacía de la agencia*— surge cuando alguna entidad humana ("el director") designa a otra ("el agente") para actuar en interés de la primera. Este tipo de problema de agencia ha sido estudiado ampliamente por economistas.[1] Se convierte en relevante para nuestra actual preocupación si las personas que crean una IA son distintas de las personas que encargan su creación. El propietario del proyecto o patrocinador (que podría ser cualquier cosa, desde una sola persona a la humanidad en su conjunto) pueden preocuparse de que los científicos y programadores que ejecutan el proyecto no vayan a actuar en beneficio del patrocinador.[2] Aunque este tipo de problema de agencia podría plantearle retos significativos al patrocinador del proyecto, no es un problema exclusivo de amplificación de inteligencia o de proyectos de IA. Problemas sobre la primacía de la agencia de este tipo son muy abundantes en las interacciones económicas y políticas humanas, y hay muchas maneras de abordarlos. Por ejemplo, el riesgo de que un empleado desleal sabotee o subvierta el proyecto podría minimizarse a través de cuidadosas comprobaciones de antecedentes del personal clave, el uso de un buen sistema control de versión para proyectos de software, y la supervisión intensiva por parte de múltiples monitores y auditores independientes. Por supuesto, este tipo de garantías tienen un coste —expanden las necesidades de personal, complican la selección de personal, obstaculizan la creatividad y sofocan el pensamiento independiente y crítico, todo lo cual podría reducir el ritmo de progreso. Estos costes podrían ser importantes, sobre todo para los proyectos que cuentan con presupuestos ajustados, o que se ven a sí mismos insertos en una carrera muy reñida en la que el ganador se lo lleva todo. En tales situaciones, los proyectos pueden escatimar en garantías procesales, creando posibilidades de que ocurran fallos en la primacía de la agencia potencialmente catastróficos.

La otra parte del problema de control es más específica del contexto de una explosión de inteligencia. Éste es el problema al que un proyecto se enfrenta cuando trata de asegurarse de que la superinteligencia que está construyendo no irá contra los intereses del proyecto. Esta parte también se puede considerar como un problema de primacía de la agencia, el *segundo problema de primacía de la*

agencia. En este caso, el agente no es un agente humano que opera en nombre de un director humano. En su lugar, el agente es el sistema superinteligente. Considerando que el primer problema de primacía de la agencia se produce sobre todo en la fase de desarrollo, el segundo problema de agencia amenaza con causar problemas sobre todo en la fase operativa de la superinteligencia.

Muestra 1 Problemas de agencia doble

El primer problema de primacía de la agencia

- Humano vs. Humano (Patrocinador ^ desarrollador).
- Se produce principalmente en fase de desarrollo.
- Se aplican técnicas de gestión estándar

El segundo problema de primacía de la agencia ("el problema del control")

- Humano vs. Superinteligencia (Proyecto ^ Sistema).
- Se produce principalmente en fase operativa (y de impulso).
- Se necesitan nuevas técnicas.

Este segundo problema de agencia plantea un desafío sin precedentes. Resolverlo requerirá nuevas técnicas. Ya hemos considerado algunas de las dificultades que entraña. Vimos, en particular, que el síndrome de giro traicionero echa a perder lo que de otro modo podría haber parecido un conjunto prometedor de métodos, los basados en la observación de la conducta de la IA en su fase de desarrollo y en la posibilidad de que la IA creciera en un entorno seguro, una vez que hubiera acumulado un historial de acciones apropiadas. La seguridad de otras tecnologías a menudo pueden evaluarse en el laboratorio o en pequeños estudios de campo, y ser luego gradualmente implementadas con la posibilidad de detener el despliegue si surgen problemas inesperados. Su rendimiento en los ensayos preliminares nos ayuda a hacer inferencias razonables sobre su fiabilidad futura. Tales métodos conductuales son inefectivos en el caso de una superinteligencia debido a la capacidad de planificación estratégica de una inteligencia general.[3]

Dado que el enfoque conductista es inútil, hay que buscar alternativas. Podemos dividir los métodos de control posibles en dos grandes clases: los *métodos de control de la capacidad,* cuyo objetivo es controlar lo que la superinteligencia puede hacer; y los *métodos de selección de la motivación,* cuyo objetivo es controlar lo que la superinteligencia quiere hacer. Algunos métodos son compatibles mientras que otros representan alternativas mutuamente excluyentes. En este capítulo sondearemos las opciones principales. (En los siguientes cuatro capítulos, vamos a explorar algunas de las cuestiones clave en mayor profundidad).

Es importante darse cuenta de que debe implementarse algún método de control (o combinación de métodos) *antes* de que un sistema llegue a ser superinteligente. Es algo que no se podrá hacer después de que el sistema haya obtenido una ventaja estratégica

decisiva. La necesidad de resolver el problema de control con antelación —y de tener éxito al poner en práctica dicha solución la primera vez que un sistema alcance la superinteligencia— es lo que hace que conseguir una detonación controlada sea un desafío de enormes proporciones.

Métodos de control de la capacidad

Los métodos de control de capacidad buscan evitar resultados indeseables limitando lo que la superinteligencia puede hacer. Esto puede implicar situar a la superinteligen- cia en un entorno en el que no sea capaz de causar daño *(métodos de encajamiento)* o en el que existan razones instrumentales fuertemente convergentes para no involucrarse en comportamientos perjudiciales *(métodos de incentivos)*. También podría implicar la limitación de las capacidades internas de la superinteligencia *(atrofia)*. Además, los métodos de control de la capacidad podrían implicar el uso de mecanismos para detectar automáticamente y reaccionar ante diversos tipos de fallo de contención o intento de transgresión *(cables trampa)*.

Métodos de encajamiento

Los métodos de encajamiento se pueden subdividir en métodos de contención física e informativa.

La contención física pretende limitar el sistema a una "caja", es decir, evitar que el sistema interaccione con el mundo externo si no es a través de los canales de salida restringidos de manera específica. El sistema encajado no tendría acceso a manipuladores físicos fuera de la caja. Extraer manipuladores (como brazos robóticos) de dentro de la caja evitaría que el sistema construyera dispositivos físicos que pudieran vulnerar el confinamiento.

Para mayor seguridad, el sistema debe ser colocado en una malla metálica para evitar la transmisión de señales de radio, que de otro modo podrían ofrecer un medio de manipulación de objetos electrónicos, tales como receptores de radio del entorno. Téngase en cuenta, por cierto, lo fácil que hubiera sido pasar por alto la necesidad de esta precaución. Uno podría ingenuamente haber asumido que un agente sin un manipulador no podía afectar al mundo externo. Pero sería posible que una inteligencia artificial generara ondas de radio, incluso cuando no tuviera acceso a manipuladores externos, simplemente "pensando" (es decir, moviendo los electrones de su circuito en patrones particulares).[4] Una vez señalado esto, la

vulnerabilidad puede ser remediada encerrando al sistema en una jaula de Faraday —pero nos quedaríamos preguntándonos cuántas otras vulnerabilidades igualmente sutiles podrían existir. Cada vez que oímos hablar de un diseño de seguridad aparentemente infalible que tiene una falla inesperada, debemos aguzar nuestros oídos. Estas ocasiones nos dan la oportunidad de abandonar una vida de exceso de confianza y resolvernos a ser mejores bayesianos.[5]

El confinamiento físico tiene varias ventajas. Es fácil de implementar. Se puede aplicar a muchas arquitecturas de inteligencia artificial, incluso a las que no fueron diseñadas inicialmente con la seguridad como un objetivo. Se puede utilizar en combinación con la mayoría de los otros métodos de control. Y parece poco probable que se vuelva en contra nuestro: es decir, aunque podría dejar de garantizar la seguridad, es poco probable que cause una catástrofe que de otro modo no se hubiera producido. La principal desventaja del confinamiento físico es que reduce la funcionalidad de la superinteligencia (un tema sobre el que volveremos en el capítulo 10). Otra preocupación es que podría alentar una falsa sensación de seguridad, aunque esto se puede evitar si consideramos el confinamiento físico como la guinda del pastel en lugar de como el elemento de precaución principal.

La contención informativa pretende restringir la información que puede salir de la caja. Ya hemos visto cómo una superinteligencia que tuviera acceso a un puerto de internet, de manera que pudiera mensajear a otras entidades externas, es potencialmente insegura: incluso si comenzara sin acceso a mecanismos físicos, podría utilizar su canal de salida de información para conseguir que los seres humanos hicieran su voluntad. Un método de contención informativo obvio, por lo tanto, es prevenir al sistema de que acceda a redes de comunicación.

Desconectar una IA de internet no consigue, sin embargo, garantizar la seguridad si hay uno o más seres humanos que sirven como guardianes del sistema y permanecen en comunicación con ella. Durante la fase de desarrollo, los programadores sirven como guardianes; durante la fase operativa, el dueño del sistema podría desempeñar este papel. Si la IA pudiera persuadir o engañar a un guardián para que le sacara fuera de la caja, y esto resulta en su acceso, ya sea a internet o directamente a mecanismos físicos, entonces la estrategia de encajamiento habría fallado. Los seres humanos no son sistemas seguros, sobre todo cuando se enfrentan a una superinteligencia astuta y persuasiva.

Aunque confiar las comunicaciones externas de una inteligencia artificial a un supervisor humano no garantice la seguridad, podría servir como una medida de seguridad auxiliar. Hay precauciones que

podrían reducir el riesgo de que la IA escapara de su guardián. Podría restringirse a la IA a comunicarse sólo a través de un interfaz de teletipo de bajo ancho de banda. Esto haría un poco más difícil de manipular psicológicamente a los porteros (evitando el uso del sonido emotivo y las imágenes y trucos tales como parpadear patrones hipnóticos para inducir un estado de trance).[7] Yendo más lejos, el contenido (no sólo el modo de presentación) de la información saliente podría ser restringido; de manera que, por ejemplo, la IA se viera limitada a dar un solo bit como respuesta a las preguntas sí / no. La idea de construir un sistema que respondiera a preguntas será examinada en el próximo capítulo.

El caso límite del enfoque de encajamiento sería un sistema mantenido en aislamiento físico e informativo. Incluso si esto fuera alcanzable, sin embargo, un sistema de este tipo aislado sería bastante inútil ya que no tendría ningún efecto sobre el mundo exterior. Podría pensarse que crear una superinteligencia y mantenerla aislada quizá tenga algún propósito científico: mediante el estudio del modelo de un sistema autónomo, se puede aprender acerca de su dinámica interna y sus patrones de comportamiento. Pero esto sería un error. Tan pronto como se observe un sistema, éste deja de estar aislado informativamente.[8] La información puede transmitirse no sólo a través de los mensajes que una IA envía a través de un "canal de salida" designado, sino también a través de cualquier observación que un extraño haga de cualquier consecuencia causal del funcionamiento de la IA, —su consumo de energía, su uso de la CPU y su uso de memoria, sus estados computacionales, o cualquier rastro dejado atrás después de que se haya apagado. Una IA que anticipara que podría ser observada en cualquiera de estas formas podría adoptar comportamientos estratégicamente diseñados para influir en los hipotéticos observadores. Por ejemplo, una IA podría "funcionar defectuosamente" de cualquier forma que aumentara la probabilidad de que, en última instancia, fuera reactivada o de que otra superinteligencia fuera creada con objetivos similares a los suyos.

Métodos de incentivos

Los métodos de incentivos implican situar al agente en un entorno en el que encuentre razones instrumentales para actuar de manera favorable a los intereses del agente principal.

Considérese la posibilidad de un multimillonario que utiliza su fortuna para crear una gran fundación de caridad. Una vez creada, la fundación podría ser poderosa —más poderosa que la mayoría de personas, incluyendo su fundador, que podrían haber donado la mayor parte de su riqueza. Para controlar la fundación, el fundador establece

su propósito en la constitución y los estatutos, y nombra un consejo de administración que simpatiza con su causa. Estas medidas constituyen una forma de selección de motivación, ya que su objetivo es dar forma a las preferencias de la fundación. Pero incluso si esos intentos de personalizar el funcionamiento interno de la organización fallaran, el comportamiento de la base quedaría circunscrito por su entorno social y legal. La fundación tendría un incentivo para obedecer la ley, por ejemplo, para que no fuera cerrada o multada. Sería un incentivo ofrecer a sus empleados una paga y condiciones de trabajo aceptable, y satisfacer las partes externas interesadas. Cualesquiera que fueran sus objetivos finales, la fundación tendría de este modo razones instrumentales para conformar su conducta según diversas normas sociales.

¿No podríamos esperar que una superinteligencia artificial estuviera igualmente obligada por la necesidad de llevarse bien con los demás actores con los que comparte escenario? Aunque esto podría parecer una manera sencilla de resolver el problema del control, no está libre de obstáculos. En particular, presupone un equilibrio de poder: las sanciones legales o económicas no pueden contener a un agente que tenga una ventaja estratégica decisiva. La integración social no puede, por tanto, ser considerada como un método de control en los escenarios de despegue rápido o moderado que tengan una dinámica en la que el ganador se lo lleva todo.

¿Qué sucedería en escenarios multipolares, en el que varias agencias emergieran tras la transición con niveles comparables de capacidad? A menos que la trayectoria predeterminada sea de despegue lento, el logro de una distribución de potencia de este tipo podría requerir orquestar cuidadosamente un ascenso en el que diferentes proyectos estuvieran sincronizados deliberadamente para evitar que cualquiera de ellos llegara a destacarse sobre los demás.[9] Incluso si el resultado fuera un escenario multipolar, la integración social no es una solución perfecta. Al confiar en la integración social para resolver el problema del control, el agente principal corre el riesgo de sacrificar una gran parte de su potencial influencia. Aunque un equilibrio de poder pudiera impedir que una IA particular asumiera el control del mundo, la IA todavía tendría algo de poder para afectar los resultados; y si ese poder se utilizara para promover algún objetivo arbitrario —maximizar la producción de clips— es probable que no se utilizara para promover los intereses del agente principal. Imaginemos que nuestro multimillonario creador de una nueva fundación permitiera que su misión fuera establecida por un generador de palabras al azar: no sería una amenaza a nivel de la especie, pero sería, sin duda, una oportunidad perdida.

Una idea relacionada pero diferente de manera relevante es que

una IA, al interactuar libremente en la sociedad, adquiriría nuevos objetivos finales de tipo humano. Tal proceso de socialización es el que tiene lugar entre nosotros los seres humanos. Interiorizamos normas e ideologías, y llegamos a valorar otros individuos en sí mismos como consecuencia de nuestras experiencias con ellos. Pero esto no es una dinámica universalmente presente en todos los sistemas inteligentes. Como se señaló anteriormente, muchos tipos de agentes en diferentes situaciones tendrían razones instrumentales convergentes para *no* permitir cambios en sus objetivos finales. (Se podría considerar la posibilidad de diseñar un tipo especial de sistema de objetivos que pudiera adquirir objetivos finales al modo en que lo hacen los humanos, pero esto no contaría como un método de control de capacidad. Discutiremos algunos de los posibles métodos de adquisición de valores en el capítulo 12).

El control de capacidad a través de la integración social y el equilibrio de poder se basa en fuerzas sociales difusas que gratifican y penalizan a la IA. Otro tipo de método de incentivos implicaría la creación de una configuración en la que la IA pudiera ser recompensada y penalizada por el proyecto que la crea, y, por lo tanto, estuviera incentivada para actuar en interés del agente principal. Para lograr esto, la IA sería colocada en un contexto de vigilancia que permitiera monitorizar o evaluar su comportamiento, ya fuera manualmente o mediante algún proceso automatizado. La IA sabría que una evaluación positiva lograría parte del resultado que desea y que una evaluación negativa dejaría de hacerlo. En teoría, la recompensa podría ayudar con el cumplimiento de algún objetivo fundamental convergente. Sin saber nada específico sobre el sistema de motivación de la IA, sin embargo, podría ser difícil calibrar el mecanismo de recompensa. Por ejemplo, podríamos terminar con una IA que estuviera dispuesta a tomar riesgos extremos por una pequeña posibilidad de obtener finalmente el control de una gran parte del universo. Podría ser costoso ofrecer a la IA una mayor utilidad esperada como recompensa por la cooperación que lo que la IA podía esperar alcanzar negándose y tratando de escapar.[10]

Una mejor alternativa podría ser la de combinar el método de incentivos con el uso de la selección de motivación para dar a la IA una meta final que la hiciera más fácil de controlar. Supongamos que una IA fuera diseñada para tener como objetivo final que un botón rojo en particular dentro de un búnker de mando nunca fuera presionado. Dado que presionar el botón es valorado negativamente de manera intrínseca y no a causa de sus consecuencias causales, el botón puede ser completamente inerte: podría estar hecho de plastilina. Además, sería irrelevante que la IA nunca pudiera saber si el botón ha sido presionado. Lo esencial es que la IA crea que es más

probable que el botón no sea pulsado si la IA actúa continuamente en interés del agente principal, en lugar de rebelarse.

Es posible matizar esta configuración. En lugar de tratar de dotar a una IA con un objetivo final que se refiera a un botón físico, se podría construir una IA que diera valor final a recibir una corriente de "fichas criptográficas de recompensa".[11] Éstas serían secuencias de números que servirían como claves para sistemas de cifrado que se habrían generado antes de que la IA fuera creada y que habrían sido introducidas en su sistema de motivación.[12] Estas secuencias de números especiales serían muy deseables para la IA, pues constituirían un tipo especial de fichas de recompensa que la IA no podría alcanzar aunque realizara un pinchazo cerebral.[13] Las claves se almacenarían en un lugar seguro donde pudieran ser destruidas rápidamente si la IA en algún momento hiciera un intento de apoderarse de ellas. En tanto que la IA coopere, las fichas se repartirían a un ritmo constante.

La IA podría entonces dar una función de utilidad limitada, y los parámetros podrían configurarse para que la IA obtuviera (por ejemplo) el 99% de utilidad máxima por la primera ficha de recompensa; el 99% de utilidad potencial restante por la segunda ficha de recompensa; etc. Al mantener una alta "tasa de descuento", y mediante la transmisión de señales de recompensa a la IA a un ritmo razonablemente alto, se podría disuadir a la IA de que intentara llevar a cabo planes a largo plazo que aumentaran el número esperado de fichas de recompensa en el futuro lejano a expensas de cierto riesgo de que su flujo de recompensa fuera interrumpido o de que la siguiente ficha de recompensa fuera cancelada y destruida. Si la IA siempre creyera que desertar (es decir, dejar de hacer lo mejor para preservar los intereses del agente principal) incurriría en al menos un 2% de riesgo de perder la siguiente ficha de recompensa, mientras que la cooperación redujera el riesgo de perder la siguiente ficha de recompensa por debajo del 1%, una IA que maximizara la utilidad de sus acciones siempre cooperaría.

¿Qué podría salir mal con tal sistema de incentivos? Una posibilidad es que la IA no confiara en que el operador humano le fuera a entregar las recompensas prometidas. El historial de fiabilidad humana es muy distinto a una línea recta de perfección infalible. La IA podría preocuparse razonablemente de que el operador cambiara de parecer, o que aumentara su nivel de desempeño, o que dejara de reconocer cuando la IA había hecho su parte. La IA también se preocuparía de que el operador pudiera quedar incapacitado. El riesgo combinado de estos fallos puede superar el riesgo de intentar hacerse con el control del mecanismo de recompensa. Incluso una IA encajada que poseyera la panoplia de superpoderes sería una fuerza poderosa. (Para una IA que desde el principio *no* estuviera en una caja, el

secuestro del mecanismo de recompensa dirigido por humanos puede ser como quitarle un caramelo a un bebé).

Otro problema con el sistema de incentivos es que presupone que podremos saber si los resultados producidos por la IA serán de nuestro interés. Como capítulos posteriores elaborarán, este presupuesto no es inocuo.

Una evaluación completa de la viabilidad de los métodos de incentivos también tendría que tener en cuenta otra serie de factores, incluyendo algunas consideraciones esotéricas que podrían hacer que tales métodos fueran más viables de lo que un análisis preliminar sugeriría. En particular, la IA puede enfrentarse a una incertidumbre

Cuadro 8. *Continúa*

Por el contrario, considérese una IA que tuviera un objetivo final más modesto, una que podría ser satisfecha con una pequeña cantidad de recursos, tales como el objetivo de recibir algunas fichas de recompensa criptográficas pre-producidas, o el objetivo de provocar la existencia de cuarenta y cinco clips virtuales. Tal IA no debería descontar esos mundos posibles en los que habitaría en una simulación. Una parte sustancial de la utilidad esperada total de la IA puede derivar de esos mundos posibles. La toma de decisiones de una IA con metas que son fácilmente susceptibles de ser satisfechas puede, por lo tanto —si se asigna una alta probabilidad a la hipótesis de simulación— estar dominada por consideraciones acerca de qué acciones producirían el mejor resultado si su mundo percibido fuera una simulación. Tal IA (incluso si no estuviera, de hecho, en una simulación) podría estar por tanto, fuertemente influenciada por sus creencias acerca de qué comportamientos serían recompensados en una simulación.

En particular si una IA con objetivos finales susceptibles de ser satisfechos considerara que en la mayoría de los mundos simulados que coinciden con sus observaciones sería recompensada si coopera (pero no si intenta escapar de su caja o contravenir los intereses de su creador), entonces puede que eligiera cooperar Por lo tanto, hemos encontrado que incluso una IA con una ventaja estratégica decisiva, que podría cumplir sus objetivos finales en mayor medida al apoderarse del mundo que absteniéndose de hacerlo, sin embargo, se resistiría a hacerlo.

> Así, ¡oh conciencia!, de
> nosotros todos haces
> unos cobardes, y la
> ardiente resolución
> original decae al *pálido*
> *mirar del pensamiento.*
> Así también enérgicas
> empresas, de
> trascendencia
> inmensa, a esa mirada
> torcieron rumbo, y sin
> acción murieron.
>
> (Shakespeare, *Hamlet,* acto III, escena 1)

Una simple línea en la arena, respaldada por la influencia de un simulador que no existe, podría resultar ser una restricción más fuerte que una puerta de acero sólido de un metro de espesor[17]

Atrofia

Otro posible método de control de la capacidad es el de limitar las facultades intelectuales del sistema o su acceso a información. Esto podría hacerse mediante la ejecución de la IA en hardware que fuera lento o de corta memoria. En el caso de un sistema encajado, la

información interna también podría ser restringida.

Atrofiar una IA de esta manera limitaría su utilidad. Así, este método se enfrenta a un dilema: demasiada poca atrofia y la IA podría tener suficiente inteligencia para buscar una manera de hacerse más inteligente (y de ahí a la dominación del mundo); demasiada atrofia y la IA no sería más que otra pieza de software tonto. Una IA radicalmente atrofiada es, sin duda, segura, pero no resolvería el problema de cómo lograr una detonación controlada: una explosión de inteligencia seguirá siendo posible y podría simplemente ser desencadenada por algún otro sistema en su lugar, quizás en una fecha un poco posterior.

Uno podría pensar que sería seguro construir una superinteligencia siempre que sólo se la diera datos sobre algún ámbito restringido de hechos. Por ejemplo, se podría crear una IA que careciera de sensores y en la que sólo se hubieran cargado previamente en su memoria datos de ingeniería petrolífera o de química de péptidos. Pero si la IA es superinteligente —si tiene un nivel sobrehumano de inteligencia *general*— tal privación general de datos no garantiza la seguridad.

Hay varias razones para ello. En primer lugar, la noción de que la información es "sobre" un tema determinado es generalmente problemática. Cualquier pieza de información en principio puede ser relevante para cualquier tema, dependiendo de la información de fondo de un determinado razonador.[18] Por otra parte, un conjunto de datos dado contiene información no sólo sobre el ámbito del que se recogieron los datos, sino también acerca de varios hechos circunstanciales. Una mente sagaz que escrutara una base de conocimientos que nominalmente versara sobre química de péptidos podría inferir cosas sobre una amplia gama de temas. El hecho de que cierta información se incluya y otra información no podía darle pistas a una IA sobre el estado de la ciencia humana, los métodos y los instrumentos disponibles para estudiar péptidos, las tecnologías de fabricación utilizadas para hacer estos instrumentos y la naturaleza de los cerebros y las sociedades que conciben dichos estudios e instrumentos. Podría ser que una *superinteligencia* pudiera conjeturar correctamente mucho a partir de lo que, para mentes humanas obtusas, sólo parecerían exiguos restos de información. Incluso sin ninguna base de conocimientos concretos, una mente suficientemente superior podría ser capaz de aprender mucho simplemente por introspección, fijándose en el funcionamiento de su propia psique—las decisiones de diseño reflejadas en su código fuente, las características físicas de sus circuitos.[19] Quizás una superinteligencia podría incluso deducir *a priori* mucho sobre las propiedades posibles del mundo (la combinación de inferencia lógica con una probable

preferencia por mundos más simples, y algunos hechos elementales implicados en la existencia de una superinteligencia como sistema de razonamiento). Podría imaginar las consecuencias de diferentes leyes posibles de la física: qué tipo de planetas se formarían, qué tipo de vida inteligente evolucionaría, qué tipo de sociedades se desarrollarían, qué tipo de métodos para resolver el problema del control se intentaron, cómo podrían frustrarse esos métodos.[20]

Por estas razones, sería imprudente confiar en la privación de información como principal freno al poder de una superinteligencia. Sin embargo, algunas formas de atrofia, incluyendo la limitación de capacidad del hardware y la prevención del desarrollo de las facultades cognitivas específicas, podrían usarse como métodos de control auxiliares en la fase de desarrollo. Podrían ser utilizados, por ejemplo, para regular el ritmo de aumento de capacidad de una IA intermedia y para permitir que las pruebas se ejecutaran en niveles de capacidad menores a los de una superinteligencia a pleno rendimiento.

Cables trampa

Un cable trampa es un mecanismo que realiza pruebas de diagnóstico en el sistema (posiblemente sin su conocimiento) y efectúa una parada si detecta signos de actividad peligrosa. Los programadores pueden entonces examinar las grabaciones de la actividad sospechosa tranquilamente, y realizar cambios en la base de código de la IA o del entorno antes de reiniciarlo. Si encuentran que el problema está arraigado con suficientemente profundidad, podrían decidir abortar todo el proyecto.

Los cables trampa difieren de los métodos de incentivos en que no confían en que el sistema sea consciente de las consecuencias de su participación en actividades prohibidas. Los cables trampa están más estrechamente relacionados con los métodos de atrofia. Al igual que la atrofia, los cables trampa podrían ser utilizados como una salvaguardia temporal, proporcionando un cierto grado de protección durante la fase de desarrollo. En principio, los cables trampa también podrían utilizarse durante la fase operativa, en particular para un sistema encajado. Sin em-

Tabla 9. ***Diferentes tipos de cables trampa***

Tipo de cómo y en cómo	Descripción y ejemplos
Comportamiento	Podrían colocarse detectores alrededor de una IA encajada para detectar intentos de violación de la contención. Por ejemplo, los detectores podrían interceptar intentos de comunicación por radio o de acceso a los recursos computacionales internos destinados a extraerse. Podría instalarse un "puerto Ethernet hacia el limbo": una aparente conexión a internet que condujera

a un interruptor de parada.

Capacidades

Las pruebas automatizadas de capacidad podrían realizarse a intervalos frecuentes para determinar la habilidad de la IA en varios ámbitos. Si la tasa de mejora fuera inesperadamente alta o si la IA alcanzara un nivel de competencia que le acercara a la zona de peligro potencial, el sistema podría ralentizarse o ponerse en pausa, con el fin de permitir a los programadores monitorizar sus desarrollos más de cerca.

Contenidos

Los procesos internos de la IA podrían ser monitorizados en busca de actividad sospechosa. Por ejemplo, cualquier cambio en la representación de los valores finales de la IA podría desencadenar un apagado automático y una revisión. El uso de los recursos y los métodos algorítmicos de la IA también podrían ser monitorizados para detectar patrones inesperados. De manera más ambiciosa, las creencias de la IA, sus planes, subobjetivos y justificaciones pueden ser examinadas de forma continua, para asegurarse de que cumplen con las expectativas de los programadores. Este seguimiento de los estados intencionales de la IA permitiría una forma particularmente poderosa de control de contenidos: la digitalización de los procesos cognitivos de la inteligencia artificial para descubrir posibles engaños —ese momento vulnerable en el que una IA poco amistosa conforma su intención de ocultar sus verdaderas intenciones.[21]

La monitorización de contenido que requiere que los estados intencionales de la IA sean transparentes para los programadores o para un mecanismo de control automático puede no ser factible para todo tipo de arquitecturas de IA. (Algunas redes neuronales, por ejemplo, son opacas, ya que representan la información de manera integral y en formas que no coinciden necesariamente con los conceptos humanos). Esto puede ser una razón para evitar el uso de este tipo de arquitecturas.

bargo, la capacidad de los cables trampa para restringir a una superinteligencia de pleno derecho es muy dudosa, pues sería difícil para nosotros asegurarnos de que dicho agente no encontrara la manera de subvertir cualquier cable trampa ideado por el intelecto humano.

Como medida de seguridad complementaria en la fase de desarrollo, sin embargo, los cables trampa son importantes. Sería una buena práctica para los diseñadores de la IA codificar sus expectativas sobre cómo se supone que funcionará el sistema, y rodear entonces a la IA con cables trampa que se activaran si alguna de estas expectativas es violada. Para hacer las pruebas más rigurosas, podrían colocarse estratégicamente "tarros de miel" para que una IA defectuosa cometiera alguna violación fácilmente observable. Por ejemplo, si una IA hubiera sido diseñada de tal manera que se supone que no debería querer acceder a internet, podría instalarse un puerto Ethernet falso (que condujera a un interruptor de parada automática) sólo para ver si la IA intentaba usarlo. (Algunos ejemplos de cables trampa se dan en la Tabla 9).

Cabe destacar que el valor de un cable trampa depende no sólo del mecanismo en sí, sino también —de manera crítica— de cómo reacciona un proyecto cuando se activa un cable trampa. Si los programadores o administradores de proyectos, impacientes por avanzar, simplemente volvieran a encenderlo —o si lo hacen después de hacer alguna modificación de contadores para evitar que el cable trampa se dispare en la próxima ejecución del sistema— entonces no se habría ganado ninguna seguridad, incluso si el cable trampa

funcionara exactamente como se pretendía.

Métodos de selección de la motivación

Los métodos de selección de la motivación buscan evitar resultados indeseables conformando lo que la superinteligencia quiere hacer. Mediante la configuración del sistema de motivación del agente y sus objetivos finales, estos métodos podrían producir una superinteligencia que *no* quisiera explotar una ventaja estratégica decisiva de manera perjudicial. Puesto que un agente superinteligente es experto en lograr sus fines, si prefiere no hacer daño (en un sentido apropiado de "daño"), entonces tendería a no causar daño (en ese sentido de "daño").

La selección de motivación puede implicar formular explícitamente una meta o un conjunto de reglas a seguir *(especificación directa)* o configurar el sistema para que pueda descubrir un conjunto apropiado de valores por sí mismo en función de algún criterio implícito o indirectamente formulado *(normatividad indirecta)*. Una de las opciones en la selección de la motivación es tratar de construir el sistema de modo que tuviera objetivos modestos y poco ambiciosos (*domesticidad*). Una alternativa a la creación de un sistema de motivación a partir de cero es seleccionar un agente que ya cuenta con un sistema de motivación aceptable y entonces aumentar las facultades cognitivas de ese agente para que sea superinteligente, garantizando también así que el sistema de motivación no se corrompa en el proceso *(aumentación)*. Echemos un vistazo a estas dos opciones.

Especificación directa

La especificación directa es el método más sencillo para resolver el problema del control. El enfoque se presenta en dos versiones, la basada en reglas y la consecuencia- lista, y consiste en tratar de definir explícitamente un conjunto de normas o valores que harán que incluso una IA superinteligente en libertad actúe de manera segura y beneficiosa. La especificación directa, sin embargo, se enfrenta a obstáculos que pueden ser insuperables, obstáculos derivados de las dificultades para determinar qué normas o valores desearíamos que guiaran a la IA, y de las dificultades para expresar esas reglas o valores en código susceptible de ser leído por ordenadores.

El ejemplo tradicional de enfoque basado en reglas es el de las "tres leyes de la robótica", concepto formulado por el autor de ciencia ficción Isaac Asimov en un cuento publicado en 1942.[22] Las tres leyes eran: (1) Un robot no debe dañar a un ser humano o, por inacción,

permitir que un ser humano sufra daño; (2) Un robot debe obedecer las órdenes que le dan los seres humanos, excepto cuando tales órdenes entren en conflicto con la Primera Ley; (3) Un robot debe proteger su propia existencia, hasta donde esta protección no entre en conflicto con la Primera o Segunda Ley. Vergonzosamente para nuestra especie, las leyes de Asimov se mantuvieron como verdad incuestionada desde hace más de medio siglo: esto a pesar de los obvios problemas de este enfoque, algunos de los cuales se exploran en los propios escritos de Asimov (Asimov, probablemente, formuló las leyes precisamente para que pudieran fallar de manera interesante, proporcionando argumentos fértiles para sus historias).[23]

Bertrand Russell, quien pasó muchos años trabajando en los fundamentos de las matemáticas, comentó una vez que "todo es vago en un grado que no te das cuenta hasta que tratas de precisarlo".[24] El dictum de Russell se aplica con creces al enfoque de especificación directa. Consideremos, por ejemplo, cómo se podría explicar la primera ley de Asimov. ¿Significa que el robot debe reducir al mínimo la probabilidad de que cualquier humano llegue a ser perjudicado? En ese caso, las otras leyes se hacen super- fluas ya que siempre es posible que la IA tome alguna acción que pudiera tener al menos algún efecto microscópico en la probabilidad de que un ser humano llegara a ser perjudicado. ¿Cómo podría un robot equilibrar el gran riesgo de que unos pocos seres humanos llegaran a ser perjudicados frente al pequeño riesgo de que muchos seres humanos fueran dañados? ¿Cómo definimos "daño", de todos modos? ¿Cómo debe compararse el daño del dolor físico frente a los daños de fealdad arquitectónica o de injusticia social? ¿Se daña a un sádico en caso de impedirle atormentar a su víctima? ¿Cómo definimos "ser humano"? ¿Por qué no considerar la posibilidad de otros seres susceptibles de ser moralmente considerados, como los animales no humanos sentientes y las mentes digitales? Cuanto más se reflexiona, más proliferan las preguntas.

Quizás la analogía más cercana a un conjunto de reglas que podrían regir las acciones de una superinteligencia que operara en el mundo de manera global es un sistema legal. Pero los sistemas jurídicos se han desarrollado a través de un largo proceso de ensayo y error, y regulan sociedades humanas que cambian de forma relativamente lenta. Las leyes pueden revisarse cuando es necesario. Y lo que es más importante, los sistemas legales son administrados por jueces y jurados que, por lo general, se ajustan a una medida de sentido común y decencia humana que ignoran posibles interpretaciones legales lógicas que son obviamente indeseables y estaban
lejos de la intención de los legisladores. Probablemente, es

humanamente imposible formular de manera explícita un conjunto muy complejo de normas detalladas, hacer que se apliquen a un conjunto muy diverso de circunstancias, y hacerlo correctamente al primer intento.

Los problemas del enfoque consecuencialista directo son similares a los del enfoque basado en reglas directo. Esto es cierto incluso si la IA estuviera dedicada a un propósito aparentemente simple como la implementación de una versión del utilitarismo clásico. Por ejemplo, el objetivo "maximizar la expectativa del placer sobre el dolor en el balance total del mundo" puede parecer simple. Sin embargo, expresarlo en código informático implicaría, entre otras cosas, especificar cómo reconocer el placer y el dolor. Hacer esto de forma fiable podría requerir la resolución de una serie de problemas persistentes de filosofía de la mente —aunque sólo fuera para obtener una representación correcta expresada en lenguaje natural, una representación que, luego, de alguna manera, habría que traducir a lenguaje de programación.

Un pequeño error en la expresión filosófica o su traducción a código podría tener consecuencias catastróficas. Considérese una IA que tuviera el hedonismo como objetivo final, y que quisiera, por tanto, alicatar el universo con "hedonium" (materia organizada en una configuración que sea óptima para la generación de experiencia placentera). Para este fin, la IA puede producir computronium (materia organizada en una configuración que es óptima para el cálculo) y utilizarlo para implementar mentes digitales en estados de euforia. Con el fin de maximizar la eficiencia, la IA omitiría de esta implementación todas las facultades mentales que no fueran esenciales para la experimentación de placer, y explotaría todos los accesos directos de cálculo que no vicien la generación de placer de acuerdo a su definición de placer. Por ejemplo, la IA podría limitar su simulación para recompensar a los circuitos, eliminar facultades tales como la memoria, la percepción sensorial, la función ejecutiva, y el lenguaje; podría simular mentes en un nivel relativamente tosco de funcionalidad, omitiendo los procesos neuronales de nivel inferior; podría reemplazar cálculos con llamadas a una tabla de consulta comúnmente repetida; o podría poner en marcha algún tipo de acuerdo por el cual múltiples mentes compartirían la mayor parte de su maquinaria computacional subyacente (sus "bases de superveniencia" en jerga filosófica). Tales trucos podrían aumentar en gran medida la cantidad de placer produci- ble con una determinada cantidad de recursos. No está claro cómo de deseable sería esto. Además, si el criterio de la IA para determinar si un proceso físico genera placer es defectuoso, entonces las optimizaciones de la IA podrían tomar la parte por el todo: podrían descartar algo que no

fuera esencial según el criterio de la IA pero que fuera esencial de acuerdo a los criterios implícitos en nuestros valores humanos. El universo entonces estaría incorrectamente cubierto de hedonium resplandeciente pero con procesos computacionales que son inconscientes y completamente inútiles —el equivalente a un cara sonriente fotocopiada trillones y trillones de veces y estampada a lo largo de las galaxias.

Domesticidad

Un tipo especial de meta final que podría ser más susceptible a la especificación directa que los ejemplos anteriores es la meta de auto-limitación. Aunque parezca muy difícil especificar cómo podríamos querer que una superinteligencia se comportarse en el mundo *en general,* ya que esto requeriría dar cuenta de todas las ventajas y desventajas en todas las situaciones que pudieran surgir, podría ser factible especificar cómo una superinteligencia debería comportarse en una situación particular. Por lo tanto, podríamos tratar de motivar al sistema para que se limitara a actuar a pequeña escala, en un contexto estrecho, y por medio de un conjunto limitado de modos de acción. Nos referiremos a este enfoque de dar a la IA objetivos finales destinados a limitar el alcance de sus ambiciones y actividades como "domesticidad"

Por ejemplo, se podría tratar de diseñar una IA tal que funcionara como un dispositivo de pregunta-respuesta (un "oráculo", para anticipar la terminología que vamos a presentar en el próximo capítulo). Dar simplemente a la IA el objetivo final de producir respuestas máximamente precisas a cualquier pregunta planteada sería poco seguro —recuérdese la "catástrofe de la hipótesis de Riemann" que se describe en el capítulo 8. (Démonos cuenta, también, que este objetivo incentivaría que la IA tomara medidas para asegurarse de que se le hicieran preguntas fáciles). Para lograr la domesticidad, se podría tratar de definir un objetivo final que de alguna manera superara estas dificultades: tal vez una meta que combinara los desiderata de responder a las preguntas correctamente y minimizara el impacto de la IA en el mundo, excepto los resultados que fueran consecuencias incidentales de dar respuestas exactas y no manipulativas a las preguntas que se le hiciera.[26]

La especificación directa de un objetivo como la domesticidad es más probable que sea factible que la especificación directa de un objetivo más ambicioso o de un conjunto de reglas completo para operar en situaciones abiertas. Persisten, sin embargo, retos significativos. Habría que tener cuidado, por ejemplo, definiendo qué

sería para la IA "minimizar su impacto en el mundo" para garantizar que la medida del impacto de la IA coincida con nuestros propios estándares de lo que entendemos como un gran o un pequeño impacto. Una mala medida daría lugar a malos resultados. Hay también otros tipos de riesgo asociados con la construcción de un oráculo, que discutiremos más adelante.

Hay una afinidad natural entre el enfoque de domesticidad y el de contención física. Se podría tratar de "encajar" a una IA de tal manera que el sistema fuera incapaz de escapar y mantener, al mismo tiempo, un sistema de motivación de la IA que hiciera que no estuviera dispuesta a escapar incluso si encontrara la manera de hacerlo. Manteniéndose las demás cosas iguales, la existencia de múltiples mecanismos de seguridad independientes deberían condensar las probabilidades de éxito.[27]

Normatividad indirecta

Si la especificación directa parece desesperada, podríamos en su lugar intentar la normatividad indirecta. La idea básica es que en lugar de especificar directamente un estándar normativo concreto, especificamos un proceso del que derivar un estándar. Luego construimos el sistema de tal modo que esté motivado para llevar a cabo este proceso y para adoptar cualquier norma en la que el proceso desemboque.[28] Por ejemplo, el proceso podría consistir en llevar a cabo una investigación sobre el problema empírico de qué versión ideal de nosotros preferiría hacer una IA. El objetivo final dado a la IA en este ejemplo podría ser algo así como "lograr lo que nos hubiera gustado que la IA alcanzara si hubiéramos pensado en el asunto larga y trabajosamente".

Una explicación más detallada de la normatividad indirecta tendrá que esperar al capítulo 13. Allí, volveremos sobre la idea de "extrapolar nuestra voluntad" y exploraremos diversas formulaciones alterativas. La normatividad indirecta es un enfoque muy importante para la selección de motivación. Su promesa radica en el hecho de que podría permitirnos descargar en la superinteligencia gran parte del trabajo cog- nitivo difícil requerido para llevar a cabo la especificación directa de un objetivo final apropiado.

Aumentación

El último método de selección de la motivación de nuestra lista es el de aumentación. Aquí la idea es que en lugar de tratar de diseñar un sistema de motivación *de novo,* partimos de un sistema que ya tiene un

sistema de motivación aceptable, y mejoramos sus facultades cognitivas para que sea superinteligente. Si todo va bien, esto nos daría una superinteligencia con un sistema de motivación aceptable.

Este enfoque, obviamente, es inútil en el caso de una IA seminal recién creada. No obstante, la aumentación es un potencial método de selección de la motivación para otras rutas de acceso a la superinteligencia, incluyendo la emulación de cerebro, la mejora biológica, los interfaces de ordenador-cerebro, y las redes y organizaciones donde hay una posibilidad de construir el sistema a partir de un núcleo normativo (seres humanos regulares) que ya contienen una representación del valor humano.

El atractivo de la aumentación puede crecer en proporción a nuestra desesperación con los otros enfoques del problema de control. La creación de un sistema de motivación para una IA seminal que siga siendo seguro y beneficioso de manera fiable bajo una situación de auto-mejoramiento recursivo, incluso cuando el sistema se convirtiera en una superinteligencia madura, es una tarea difícil, sobre todo si tuviéramos que lograr la solución correcta en el primer intento. Con la aumentación, al menos comenzaríamos con un sistema que tendría motivaciones familiares y de apariencia humana.

En el lado negativo, puede ser difícil asegurar que un sistema de motivación complejo evolucionado, mal y torpemente comprendido, como el del ser humano, no se corrompa cuando tenga una explosión cognitiva estratosférica. Como se mencionó anteriormente, un procedimiento de emulación cerebral imperfecto que conservara el funcionamiento intelectual puede que no conservara todas las facetas de la personalidad. Lo mismo es cierto (aunque tal vez en menor medida) para las mejoras biológicas en la cognición, lo que podría afectar sutilmente a la motivación; y respecto de las mejoras en inteligencia colectiva de organizaciones y redes, que podrían cambiar de manera adversa la dinámica social (por ejemplo, en formas que lleven a una actitud denigrante hacia el colectivo de los extranjeros o hacia sus propios electores). Si se lograra la superinteligencia a través de cualquiera de estos caminos, el patrocinador del proyecto tendría difícil conseguir garantías sobre las motivaciones últimas del sistema maduro. Una arquitectura de IA matemáticamente bien especificada y fundamentalmente elegante podría ofrecer —por su carácter de otredad no-antropomórfica— mayor transparencia, tal vez incluso podría ofrecer la posibilidad de que los aspectos importantes de su funcionalidad pudieran ser formalmente verificados.

Al final, sin embargo, da igual el modo en que calculemos las ventajas y desventajas de la aumentación, pues la opción de confiar en este método podría venirnos forzada. Si la superinteligencia se lograra primero por el camino de la inteligencia artificial, la aumentación no

sería aplicable. Por el contrario, si la superinteligencia se lograra primero a lo largo de algún camino distinto al de la IA, entonces muchos de los otros métodos de selección de motivación serían inaplicables. Aun así, las opiniones sobre la probabilidad de que la aumentación llegara a tener éxito tienen relevancia estratégica en la medida en que tenemos oportunidades para influir sobre qué tecnología producirá la superinteligencia primero.

Sinopsis

Es conveniente un resumen rápido antes de cerrar este capítulo. Distinguimos dos grandes clases de métodos para tratar con el problema de la agencia en el centro de seguridad de una IA: el control de la capacidad y la selección de la motivación. En la tabla 10 encontramos un resumen.

Tabla 10. ***Métodos de control***

Control de la capacidad	
Métodos de encajonamiento	Métodos de encajonamiento de manera que puede afectar al mundo externo solamente a través de algún canal restringido y previamente aprobado. Abarca métodos de contención física e informativa.
Métodos de incentivos	Métodos de incentivos en un entorno que ofrece incentivos apropiados. Esto podría implicar la integración social en un mundo de entidades igualmente poderosas. Otra variación es el uso de fichas (criptográficas) de recompensa. "La captura antrópica" también es una posibilidad muy importante pero implica consideraciones esotéricas.
Estratificación	Se imponen restricciones a las capacidades cognitivas del sistema o a la capacidad de afectar procesos internos clave.
Señuelos	Se realizan pruebas de diagnóstico en el sistema (posiblemente sin su conocimiento) y un mecanismo apaga el sistema si detecta actividad peligrosa.
Selección de la motivación	
Especificación directa	Especificación directa de algún sistema de motivación especificado directamente, lo que podría hacerse de manera consecuencialista o mediante el seguimiento de una serie de reglas.
Domesticidad	Se diseña un sistema de motivación para limitar seriamente el alcance de las ambiciones y las actividades del agente.
Normatividad indirecta	La normatividad indirecta podría estar basada en reglas o en principios consecuencialistas, pero se distingue por su dependencia de un enfoque indirecto de especificación de reglas a seguir o de valores que han de ser perseguidos.
Aumentación	Se empieza con un sistema que ya tiene motivaciones sustancialmente humanas o benevolentes, y se mejora sus capacidades cognitivas para que llegue a ser superinteligente.

Cada método de control tiene sus posibles vulnerabilidades y presenta diferentes grados de dificultad en su ejecución. Podría tal vez pensarse que debemos clasificarlos de mejor a peor, y luego optar por el mejor método. Pero eso sería simplista. Algunos métodos se pueden utilizar de manera combinada, mientras que otros son excluyen- tes. Incluso un método comparativamente inseguro puede ser aconsejable si se

puede utilizar fácilmente como complemento, mientras que un método fuerte podría ser poco atractivo si supusiera excluir otras salvaguardas deseables.

Por lo tanto, es necesario considerar qué paquete de ofertas hay disponibles. Tenemos que considerar qué tipo de sistema podríamos tratar de construir y qué métodos de control serían aplicables a cada tipo. Este es el tema de nuestro próximo capítulo.

CAPÍTULO 10

Oráculos, genios, soberanos, herramientas

Algunos dicen: "¡Simplemente construid un sistema de pregunta-respuesta!" o "¡Simplemente construid una IA que sea como una herramienta y no como un agente!" Sin embargo, estas sugerencias no hacen que todos los problemas de seguridad se esfumen, y, de hecho, no es una cuestión trivial qué tipo de sistema ofrecería las mejores perspectivas de seguridad. Consideraremos cuatro tipos o "castas" —oráculos, genios, soberanos y herramientas— y explicaremos cómo se relacionan entre sí.[1] Cada uno ofrece diferentes conjuntos de ventajas y desventajas para nuestro propósito de resolver el problema del control.

Oráculos

Un oráculo es un sistema de pregunta-respuesta. Podría aceptar preguntas en lenguaje natural y presentar sus respuestas como texto. Un oráculo que sólo aceptara preguntas de tipo sí / no podría exteriorizar su mejor conjetura con un solo bit, o tal vez con bits adicionales para representar su grado de confianza. Un oráculo que aceptara preguntas abiertas necesitaría alguna métrica con la que clasificar las posibles respuestas veraces en cuanto a su capacidad informativa o grado de propiedad.[2] En cualquier caso, la construcción de un oráculo que tuviera una habilidad total de ámbito general para responder a preguntas en lenguaje natural es un problema de IA completa. Si se pudiera hacer eso, probablemente también se podría construir una IA que tuviera una capacidad decente para entender las intenciones y palabras humanas.

Oráculos con ámbitos limitados de superinteligencia también son concebibles. Por ejemplo, se podría concebir un oráculo matemático que sólo aceptara consultas formuladas en lenguaje formal, pero que fuera muy bueno en responder a tales preguntas (por ejemplo, que fuera capaz de resolver en un instante casi cualquier problema matemático expresado formalmente que los matemáticos humanos sólo podrían resolver trabajando colaborativamente durante un siglo). Tal oráculo matemático constituiría un trampolín hacia la superinteligencia de ámbito general.

Ya existen oráculos con superinteligencia en ámbitos muy limitados. Una calculadora de bolsillo puede ser vista como un oráculo muy reducido para preguntas aritméticas básicas; un buscador de internet puede ser visto como una realización muy parcial de un oráculo cuyo ámbito abarca una parte importante del conocimiento declarativo humano general. Estos oráculos de ámbito limitado son herramientas en vez de agentes (más sobre IAs-herramienta en breve). En lo que sigue, sin embargo, el término “oráculo” se referirá a los sistemas de pregunta-respuesta que tienen una superinteligencia de ámbito general, a menos que se indique lo contrario.

Para hacer que una superinteligencia general funcionara como un oráculo, podríamos aplicar tanto la selección de motivación como el control de capacidad. Seleccionar la motivación para un oráculo puede ser más fácil que para otras castas de superinteligencia, porque el objetivo final de un oráculo podría ser relativamente sencillo. Queremos que el oráculo dé respuestas veraces y no manipulativas, y queremos limitar cualquier otro impacto suyo en el mundo. Aplicando un método de domesticidad, podríamos obligar al oráculo a utilizar sólo los recursos designados para producir su respuesta. Por ejemplo, podríamos establecer que tuviera que basar su respuesta en un corpus precargado de información, como una base fija de internet, y que no pudiera emplear más de un número fijo de pasos computacionales.[3]

Para evitar incentivar que el oráculo nos manipule para que le demos preguntas más fáciles —algo que ocurriría si le diéramos el objetivo de maximizar su precisión en todas las preguntas que le pidiéramos— podríamos darle el objetivo de responder a una sola pregunta y poner fin inmediatamente después de dar su respuesta. Se cargaría la pregunta en su memoria antes de ejecutar el programa. Para hacer una segunda pregunta, tendríamos que reiniciar la máquina y ejecutar el mismo programa con una pregunta diferente precargada en su memoria.

Desafíos sutiles y potencialmente peligrosos surgen incluso en la especificación de sistemas de motivación relativamente simples como los necesarios para ejecutar un oráculo. Supongamos, por ejemplo, que nos encontramos con alguna explicación de lo que significa para la IA "minimizar su impacto en el mundo, sujeto a la consecución de determinados resultados" o "utilizar sólo los recursos designados en la preparación de la respuesta". ¿Qué sucede si la IA, en el curso de su desarrollo intelectual, sufre el equivalente a una revolución científica que implique un cambio en su ontología básica?[4] Inicialmente podríamos haber explicado "impacto" y "recursos designados" por medio de nuestra propia ontología (postulando la existencia de varios objetos físicos tales como los ordenadores). Pero así como hemos abandonado categorías on- tológicas que fueron dadas por supuesto por los científicos en edades anteriores (por ejemplo, "flogisto", "élan vital" y "simultaneidad absoluta"), una IA superinteligente podría descubrir que algunas de nuestras categorías actuales se basan en conceptos fundamentalmente erróneos. El sistema de objetivos de una IA que se sometiera a una crisis ontológica debería ser lo suficientemente resistente como para que el "espíritu" de su contenido objetivo original se conservara, caritativamente transpuesto al nuevo paradigma.

A pesar de que hacer un oráculo seguro mediante el uso de la selección de la motivación podría estar lejos de ser trivial, puede, sin embargo, ser más fácil que hacer lo mismo para una IA que vagara por el mundo en busca de un objetivo complicado. Esto es un argumento para preferir que la primera superinteligencia sea un oráculo. Otro punto a favor del oráculo como primer camino es la mayor susceptibilidad de un oráculo al control de capacidad. Todos los métodos de encajamiento estándar tienen aplicación aquí. Además, puede haber métodos que sean distintivamente aplicables a oráculos. Por ejemplo, considérese el riesgo de que un oráculo respondiera a las preguntas no de una manera máximamente veraz, sino de manera tal que nos manipulara sutilmente en la promoción de sus propios planes ocultos. Una forma de mitigar ligeramente esta amenaza podría ser la creación de múltiples oráculos, cada uno con un código ligeramente

diferente y con una base de información ligeramente diferente. Un sencillo mecanismo podría entonces comparar las respuestas dadas por los diferentes oráculos y sólo presentárselos a los humanos si todas las respuestas estuvieran de acuerdo. Mientras que uno debe asumir que cualquier oráculo superinte- ligente de ámbito general sabría de la existencia de otros oráculos (podría inferir esto desde su conocimiento de la sociedad humana), podría ser factible ocultar algunos detalles de implementación de cada oráculo de la vista del resto. Si los oráculos no fueran capaces de comunicarse entre sí, podría entonces ser difícil para ellos coordinarse sobre la manera de responder de manera manipuladora a nuestras preguntas. Hay muchas maneras de desviarse de la verdad, y los oráculos no estarían todos de acuerdo en cuál de estas desviaciones sería más atractiva —mientras que la verdad misma es un punto de Schelling (un lugar destacado para el acuerdo en ausencia de comunicación). Así que si los oráculos lograran consenso, podría ser una señal de que nos dieron la respuesta verdadera.[5]

Un oráculo sería idealmente digno de confianza si pudiéramos asumir con seguridad que sus respuestas son siempre exactas, en la medida de sus posibilidades. Pero incluso un oráculo no confiable podría ser útil. Podríamos pedirle a dicho oráculo preguntas para las que sea difícil encontrar la respuesta, pero fácil de verificar si una respuesta dada es correcta. Muchos de los problemas matemáticos son de este tipo. Si nos preguntamos si una proposición matemática es cierta, podríamos preguntarle al oráculo para producir una prueba o refutación de la proposición. Encontrar la prueba puede requerir conocimiento y creatividad más allá de nuestro alcance, pero la comprobación de la validez de una supuesta prueba se puede hacer por un procedimiento mecánico simple.

Si es difícil verificar respuestas (como suele ser el caso en temas más allá de la lógica y las matemáticas), podemos seleccionar aleatoriamente un subconjunto de las respuestas del oráculo para la verificación. Si son correctas, podemos asignar una alta probabilidad de que la mayoría de las otras respuestas también sean correctas. Este truco puede darnos un descuento mayor en respuestas confiables que sería costoso de verificar individualmente. (Por desgracia, no podría darnos respuestas confiables que *no fuéramos capaces* de verificar, ya que un oráculo encubridor puede optar por responder correctamente sólo aquellas preguntas que crea que seremos capaces de comprobar).

Podría haber cuestiones importantes de las que podríamos aprovecharnos si apuntáramos desde el principio a la respuesta correcta (o hacia un método para localizar la respuesta correcta), incluso aunque desconfiáramos activamente de su procedencia. Por ejemplo, se podría pedir la solución a diversos problemas técnicos o filosóficos que puedan surgir en el curso de tratar de desarrollar

métodos de selección de la motivación más avanzados. Si tuviéramos una propuesta de diseño de IA supuestamente segura, podíamos pedir a un oráculo que identificara cualquier defecto importante del diseño, y preguntarle si sería capaz de explicarnos a nosotros cualquier defecto en veinte palabras o menos. Preguntas de este tipo podrían obtener información valiosa. Se requeriría precaución y moderación, sin embargo, para que no hiciéramos *demasiadas* preguntas de ese tipo —y no confiarnos *demasiado* de los detalles de las respuestas dadas a las preguntas realizadas— no sea que le demos oportunidades a un oráculo de poca confianza de que nos coma la cabeza (mediante mensajes aparentemente plausibles, pero de sutil manipulación). Puede que no sean necesarios muchos bits de comunicación para que una IA con el superpoder de la manipulación social nos doblegue a su voluntad.

Incluso si el propio oráculo funciona exactamente como se espera, existe el riesgo de que fuera mal utilizado. Una dimensión evidente de este problema es que una IA oráculo sería una fuente de inmenso poder que podría dar una ventaja estratégica decisiva a su operador. Este poder puede ser ilegítimo y puede no ser utilizado para el bien común. Otra dimensión más sutil pero no menos importante es que el uso de un oráculo podría ser extremadamente peligroso para el operador mismo. Preocupaciones similares (que implican tanto cuestiones filosóficas como técnicas) surgen también para otras castas hipotéticas de superinteligencia. Las exploraremos más a fondo en el capítulo 13. Baste aquí señalar que el protocolo para determinar qué preguntas se hacen, en qué secuencia, y cómo se comunican y difunden las respuestas, podría ser de gran importancia. También se podría considerar la posibilidad de tratar de construir un oráculo de tal manera que se negara a responder a cualquier pregunta de la que predijera una respuesta que pudiera contener consecuencias clasificadas como catastróficas según ciertos criterios improvisados.

Genios y soberanos

Un genio es un sistema de ejecución de órdenes: cuando recibe una orden de alto nivel, la lleva a cabo, y hace una pausa para esperar hasta la próxima orden.[6] Un soberano es un sistema que tiene un mandato abierto para operar en el mundo en la búsqueda de objetivos generales y, posiblemente, de largo alcance. Aunque éstos pueden parecer plantillas radicalmente diferentes de lo que una superinteligencia debería ser y hacer, la diferencia no es tan profunda como podría parecer a primera vista.

Con un genio, se sacrifica la propiedad más atractiva de un

oráculo: la oportunidad de utilizar los métodos de encajamiento. Si bien se podría considerar la creación de un genio limitado físicamente, por ejemplo, uno que sólo pudiera construir objetos dentro de un designado volumen —un volumen que podría estar cerrado por una pared endurecida o una barrera cargada con cargas explosivas preparada para detonarse si la contención es sobrepasada— sería difícil tener mucha confianza en la seguridad de cualquiera de estos métodos de contención física contra una superinteli- gencia equipada con manipuladores versátiles y materiales de construcción. Incluso si hubiera algún modo posible para asegurar una contención tan segura como la que se puede conseguir para un oráculo, no está claro cuánto habríamos ganado dando acceso directo a la superinteligencia a manipuladores, comparado con pedirle, en su lugar, que nos diera unos planos que pudiéramos inspeccionar y luego usar para lograr el mismo resultado nosotros mismos. La ganancia en velocidad y las facilidades de saltarse el intermediario humano parece que no vale la pena la pérdida que supondría no poder usar los métodos de encajamiento fuerte disponibles para contener al oráculo.

Si se llegara a crear un genio, sería deseable construirlo de tal modo que obedeciera la intención detrás de la orden, más que su significado literal, ya que un genio literal (uno lo suficientemente superinteligente como para lograr una ventaja estratégica decisiva) podría tener una propensión a matar al usuario y al resto de la humanidad en su primer uso, por razones explicadas en la sección sobre los modos de fallo malignos en el capítulo 8. En términos más generales, sería importante que el genio buscara una interpretación caritativa —y que los seres humanos consideraran razonable— de lo que se le estuviera mandando, y que el genio estuviera motivado para llevar a cabo la orden bajo esa interpretación en lugar de bajo una interpretación literal. El genio ideal sería un supermayordomo en lugar de un sabio autista.

Un genio dotado de una naturaleza tan servicial, sin embargo, no estaría lejos de pertenecer a la casta de los soberanos. Consideremos, por comparación, la idea de construir un soberano con el objetivo final de obedecer el espíritu de las órdenes que le hubiéramos dado si hubiéramos construido un genio en lugar de un soberano. Tal soberano imitaría al genio. Al ser superinteligente, este soberano haría un buen trabajo en adivinar los mandatos que habríamos dado al genio (y siempre nos podría preguntar si eso ayudara a informar a sus decisiones). ¿Habría entonces realmente alguna diferencia importante entre un soberano y un genio? O, si contemplamos la distinción desde el otro lado, consideramos que un genio superinteligente podrá igualmente ser capaz de predecir las órdenes que vamos a darle: ¿qué se gana de tener que esperar propiamente la orden antes de actuar?

Uno podría pensar que una gran ventaja de un genio sobre un soberano es que si algo sale mal, podríamos dar al genio una nueva orden para detener o revertir los efectos de las acciones anteriores, mientras que un soberano simplemente seguiría adelante sin tener en cuenta nuestras protestas. Pero esta aparente ventaja en seguridad respecto del genio es en gran parte ilusoria. El botón de "stop" o el de "deshacer" en un genio sólo funciona para los modos de fallo benignos: en el caso de un fallo maligno —uno en el que, por ejemplo, llevar a cabo la orden existente se hubiera convertido en un objetivo final para el genio— el genio simplemente ignoraría cualquier intento posterior de contravenir la orden anterior.[7]

Una opción sería la de tratar de construir un genio tal que presentara automáticamente al usuario una predicción acerca de los aspectos más destacados de los posibles resultados de una orden propuesta, pidiendo confirmación antes de proceder. Este sistema podría ser referido como uno de *genio-con-vista-previa.* Pero si esto se pudiera hacer para un genio, podría igualmente hacerse para un soberano. Así que, de nuevo, esto no es un diferenciador claro entre genio y soberano. (Suponiendo que una funcionalidad de vista previa se pudiera crear, las cuestiones de si es posible, y si es así cómo usar dicha función, son bastante menos evidentes de lo que parece, no obstante el fuerte atractivo que constituiría poder echar un vistazo a los resultados antes de comprometerse con lo que será la realidad irrevocable. Volveremos sobre este asunto más adelante).

La capacidad de una casta para imitar a otra se extiende también a los oráculos. Se podría crear un genio para que actuara como un oráculo si las únicas órdenes que le diéramos consistieran en responder a ciertas preguntas. Podría crearse un oráculo, a su vez, para sustituir a un genio si le preguntáramos al oráculo cuál es la forma más fácil de conseguir que determinadas órdenes fueran ejecutadas. El oráculo podría darnos paso a paso las instrucciones para lograr el mismo resultado que lograría un genio, o incluso podría emitir el código fuente de un genio.[8] Argumentos similares pueden esgrimirse respecto a la relación entre un oráculo y un soberano.

La verdadera diferencia entre las tres castas, por lo tanto, no reside en las capacidades finales que desbloquearían. La diferencia radicaría, en cambio, en sus distintos enfoques para el problema de control. A cada casta le corresponde un conjunto diferente de precauciones de seguridad. La característica más prominente de un oráculo es que puede ser encajado. También se podría tratar de aplicar la selección de motivación de domesticidad a un oráculo. Un genio es más duro de encajar, pero al menos puede aplicársele la domesticidad. Un soberano no puede ser ni encajado ni manejado a través de la domesticidad.

Si estos fueran los únicos factores relevantes, el orden de preferencia parece claro: un oráculo sería más seguro que un genio, que a su vez sería más seguro que un soberano; y las diferencias iniciales en conveniencia y velocidad de operación serían relativamente pequeñas y fácilmente compensadas por los logros en materia de seguridad que se pueden obtener mediante la construcción de un oráculo. Sin embargo, hay otros factores que deben tenerse en cuenta. Al elegir entre las castas, se debe considerar no sólo el peligro que representa el sistema en sí, sino también los peligros que surgen de la forma en que podría ser utilizado. Un genio obviamente otorga a la persona que lo controla un enorme poder, pero lo mismo vale para un oráculo.[9] Un soberano, por el contrario, podría ser construido de tal forma que no se le concediera a ninguna persona o grupo una influencia especial sobre el resultado, de tal manera que se resistiera a cualquier intento de corromper o alterar su agenda original. Es más, si la motivación de un soberano se definiera utilizando la "normatividad indirecta" (un concepto que explicado en el capítulo 13), entonces podría ser utilizado para lograr algún resultado definido de manera abstracta, como que "todo sea máximamente justo y moralmente correcto" —sin que nadie sepa de antemano qué es exactamente lo que esto implica. Esto crearía una situación análoga al "velo de ignorancia" de Rawls.[10] Una configuración de este tipo podría facilitar la consecución de un consenso, ayudar a prevenir los conflictos y propiciar un resultado más equitativo.

Otro punto, que cuenta en contra de algunos tipos de oráculos y genios, es que hay riesgos involucrados en el diseño de superinteligencias que tengan un objetivo final que no coincida plenamente los resultados que en última instancia queremos alcanzar. Por ejemplo, si usamos una motivación de domesticidad para que la superinteli- gencia busque minimizar su impacto en el mundo, puede que de esta manera creemos un sistema cuya jerarquía entre los posibles resultados preferibles difiera de la del patrocinador. Lo mismo sucederá si construimos la IA para que dé un valor alto a responder a las preguntas correctamente u obedecer fielmente órdenes individuales. Ahora bien, si se tiene cuidado, esto no debería causar ningún problema: habría suficiente acuerdo entre las dos jerarquías —al menos en la medida en que pertenecieran a mundos posibles que tuvieran una oportunidad razonable de ser actualizados— como para que los resultados que fueran buenos según el estándar de la IA también fueran buenos según la norma del director. Pero tal vez podría argumentarse, en relación al principio de diseño, que no es prudente introducir siquiera una cantidad limitada de falta de armonía entre nuestras metas y las de la IA. (La misma preocupación, por supuesto, se aplica a dar a los soberanos metas que no armonizan

completamente con las nuestras).

IAs-herramienta

Una sugerencia que se ha hecho es que construyamos la superinteligencia para que sea como una herramienta y no como un agente.[11] Esta idea parece surgir de la observación de que el software ordinario, que se utiliza en innumerables aplicaciones, no plantea problemas de seguridad ni remotamente análogos a los desafíos que se plantean en este libro. ¿Por qué no crear una "IA-herramienta" que fuera igual que el software —como un sistema de control de vuelo, por ejemplo, o un asistente virtual— sólo que más flexible y capaz? ¿Por qué construir una superinteligencia que tuviera una voluntad propia? En esta línea de pensamiento, el paradigma del agente es fundamentalmente erróneo. En lugar de crear una IA que tuviera creencias y deseos y que actuara como una persona jurídica, debemos tratar de construir software regular que simplemente hiciera lo que estuviera programado para hacer.

Esta idea de la creación de software que "simplemente hace lo que está programado para hacer" no es, sin embargo, tan sencilla si el producto que se está creando es una poderosa inteligencia general. Hay, por supuesto, un sentido trivial en el que todo el software simplemente hace lo que está programado para hacer: el comportamiento se especifica matemáticamente mediante un código. Pero esto es igualmente cierto para todas las castas de la inteligencia artificial, sean "IA-herramienta" o no. Si, en cambio, "simplemente hacer lo que está programado para hacer" significa que el software se comporta como los programadores *querían,* hay que decir que éste es un estándar que el software ordinario no cumple muy a menudo.

Debido a las capacidades limitadas del software contemporáneo (en comparación con los de la superinteligencia artificial) las consecuencias de tales fracasos son manejables, yendo desde insignificantes hasta muy costosas, pero en ningún caso suponiendo un amenaza existencial.[12] Sin embargo, si es la capacidad insuficiente en lugar de la fiabilidad suficiente lo que hace que el software ordinario actual sea exis- tencialmente seguro, entonces no está claro cómo este tipo de software podría ser un modelo para una superinteligencia segura. Se podría pensar que al ampliar la gama de tareas realizadas por el software ordinario, se podría eliminar la necesidad de una inteligencia general artificial. Sin embargo, el alcance y la diversidad de tareas que una inteligencia general podría realizar de manera rentable en una economía moderna sería enorme. Sería inviable crear software de propósito especial para manejar todas esas tareas. Incluso

si se pudiera hacer, tal proyecto tomaría mucho tiempo en llevarse a cabo. Antes de que pudiera completarse, la naturaleza de algunas de las tareas habrían cambiado, y nuevas tareas se habrían vuelto relevantes. Sería una gran ventaja contar con un software que pudiera aprender por sí solo a hacer nuevas tareas, y que descubriera nuevas tareas que necesitaran ser hechas. Pero esto requeriría que el software fuera capaz de aprender, razonar y planificar, y hacerlo de una manera potente y robusta en ámbitos cruzados. En otras palabras, se requeriría una inteligencia artificial general.

El desarrollo de software en sí mismo es una tarea especialmente relevante para nuestros propósitos. Habría enormes ventajas prácticas en ser capaz de automatizarlo. Sin embargo, la capacidad de auto-mejora rápida es justo la propiedad crítica que permitiría a una IA seminal poner en marcha una explosión de inteligencia.

Si la inteligencia general no es prescindible, ¿hay alguna otra manera de interpretar la idea de IA-herramienta a fin de conservar los caracteres tranquilizadores y pasivos de una herramienta rutinaria? ¿Se podría tener una inteligencia general que no fuera un agente? Intuitivamente, no es sólo la capacidad limitada de software común lo que hace que sea seguro, sino también su falta de ambición. No hay subrutina en Excel que secretamente quisiera dominar el mundo si fuera lo suficientemente inteligente como para encontrar la manera. La aplicación de la hoja de cálculo no "quiere" nada en absoluto; ejecuta ciegamente las instrucciones del programa. ¿Qué (uno podría preguntarse) se interpone en el camino de la creación de una aplicación más generalmente inteligente del mismo tipo? Un oráculo, por ejemplo, que, cuando se le solicitara la descripción de una meta, respondiera con un plan para lograrlo, de la misma forma en que Excel responde a una columna de números mediante el cálculo de una suma —sin expresar así ninguna "preferencia" sobre a su resultado o sobre cómo los seres humanos podrían optar por usarlo.

La forma clásica de escritura de software requiere que el programador entienda la tarea a realizar con suficiente detalle como para formular un proceso de solución explícito que consista en una secuencia de pasos matemáticamente bien definidos expresables en código.[13] (En la práctica, los ingenieros de software confían en las bibliotecas de código con comportamientos útiles recopilados, que pueden invocar sin necesidad de entender cómo se implementan los comportamientos. Pero ese código fue creado originalmente por programadores que tenían un conocimiento detallado de lo que estaban haciendo). Este enfoque funciona para resolver tareas bien entendidas, y lo acredita la mayoría del software que está actualmente en uso. No funciona, sin embargo, cuando nadie sabe con precisión la forma de resolver todas las tareas que deben llevarse a cabo. Aquí es

donde las técnicas del campo de la inteligencia artificial se vuelven relevantes. En aplicaciones reducidas, el aprendizaje automático puede ser utilizado simplemente para ajustar algunos parámetros en un programa mayormente diseñado por humanos. Un filtro de correo no deseado, por ejemplo, podría ser entrenado en un corpus de mensajes de correo electrónico seleccionados a mano en un proceso que cambie el peso dado a diversas funciones de diagnóstico en los algoritmos de clasificación. En una aplicación más ambiciosa, el clasificador puede construirse para que pueda descubrir nuevas funciones por sí mismo y poner a prueba su validez en un entorno cambiante. Un filtro de correo no deseado aún más sofisticado podría estar dotado de cierta capacidad para razonar acerca de las ventajas y desventajas a las que se enfrenta el usuario o acerca de los contenidos de los mensajes que está clasificando. En ninguno de estos casos se necesita que el programador sepa la mejor manera de distinguir la basura del jamón, solamente cómo configurar un algoritmo que pueda mejorar su propio desempeño a través del aprendizaje, el descubrimiento, o el razonamiento.

Con los avances en inteligencia artificial, sería posible para el programador descargar más del trabajo cognitivo necesario para encontrar la manera de realizar una tarea determinada. En un caso extremo, el programador simplemente especificaría un criterio formal de lo que cuenta como éxito y dejaría en manos de la IA encontrar una solución. Para guiar su búsqueda, la IA utilizaría un conjunto de heurísticas poderosas y otros métodos para descubrir la estructura de posibles soluciones. Seguiría buscando hasta encontrar una solución que satisficiera el criterio de éxito. La IA entonces aplicaría o bien la propia solución o (en el caso de un oráculo) daría la solución al usuario.

Formas rudimentarias de este enfoque son utilizadas hoy con bastante amplitud. Sin embargo, el software que utiliza técnicas de IA y de aprendizaje automático, aunque tiene cierta capacidad para encontrar soluciones que los programadores no habían anticipado, funciona a todos los efectos prácticos como una herramienta y no supone ningún riesgo existencial. Entraríamos en la zona de peligro sólo cuando los métodos utilizados en la búsqueda de soluciones se volvieran extremadamente potentes y generales: es decir, cuando empezaran a ascender a un nivel general de inteligencia y sobre todo cuando empezaran a ascender hasta la superinteligencia.

Hay (al menos) dos lugares donde podrían entonces surgir problemas. En primer lugar, el proceso de búsqueda superinteligente podría encontrar una solución que no sólo fuera inesperada sino radicalmente involuntaria. Esto podría conducir a un tipo fallo de los discutidos anteriormente (“suplantación perversa”, “profusión

infraestruc- tural" o "crimen mental"). Es más evidente cómo esto podría suceder en el caso de un soberano o de un genio, que implementan directamente la solución que han encontrado. Si hacer caras sonrientes moleculares o transformar el planeta en clips es la primera idea que la superinteligencia descubre que cumple el criterio de solución, entonces tendremos caras sonrientes o clips.[14] Pero incluso un oráculo, que —si todo va bien— simplemente *informe* de la solución, podría convertirse en la causa de una suplantación perversa. El usuario solicita al oráculo un plan para lograr un determinado resultado, o para que una tecnología que cumpla una determinada función; y cuando el usuario sigue el plan o construye la tecnología, una suplantación perversa sobreviene, como si la IA hubiera implementado ella misma la solución.[15]

Un segundo lugar en el que podrían surgir problemas es en la operatividad del software. Si los métodos que el software utiliza para buscar una solución son lo suficientemente sofisticados, pueden incluir disposiciones para gestionar el proceso de búsqueda de manera inteligente. En este caso, la máquina que ejecuta el software puede comenzar a parecer menos una mera herramienta y más un agente. Por lo tanto, el software puede empezar desarrollando un plan en busca de una solución. El plan puede especificar qué áreas explorar primero y con qué métodos, qué datos reunir, y cómo hacer mejor uso de los recursos computacionales disponibles. En la búsqueda de un plan que satisfaga el criterio interno del software (como dar una probabilidad suficientemente alta de encontrar una solución que satisfaga el criterio especificado por el usuario dentro del tiempo asignado), el software puede tropezar con una idea poco ortodoxa. Por ejemplo, se podría generar un plan que comenzara con la adquisición de recursos computacionales adicionales y la eliminación de potenciales elementos disruptores (tales como los seres humanos). Tales planes "creativos" salen a la luz cuando las habilidades cognitivas del software alcanzan un nivel suficientemente alto. Cuando el software pone su plan en acción, una catástrofe existencial puede sobrevenir.

Como los ejemplos ilustrados en el cuadro 9, los procesos de búsqueda abierta en ocasiones dan lugar a extrañas e inesperadas soluciones no antropocéntricas, incluso en sus formas actualmente limitadas. Los procesos de búsqueda de hoy en día no son peligrosos, ya que son demasiado débiles para descubrir el tipo de plan que podría permitir a un programa dominar el mundo. Este plan incluiría medidas extremadamente difíciles, como la invención de una nueva tecnología de armamento varias generaciones por delante del estado actual de la técnica, o la ejecución de una campaña de propaganda mucho más eficaz que cualquier comunicación ideada por demagogos

humanos. Para tener la oportunidad siquiera de *concebir* tales ideas, y mucho menos desarrollarlas de una manera que realmente funcione, una máquina probablemente tendría que tener la capacidad de representarse el mundo de una manera que fuera al menos tan rica y realista como el modelo del mundo poseído por un adulto humano normal (aunque la falta de conciencia en algunas áreas, posiblemente, podría ser compensada por una ha-

Cuadro 9. *Continúa*

Por ejemplo, antes de la década de 1960, era al parecer bastante común que los biólogos sostuvieran que las poblaciones de depredadores restringen su propia descendencia con el fin de evitar caer en una trampa maltusiana.[17] Aunque la selección individual trabajaría en contra de tal restricción, en ocasiones se pensaba que la selección de grupo superaría los incentivos individuales para explotar las oportunidades de reproducción y favorecerían los rasgos que beneficiaran al grupo o a la población en general. Estudios de análisis y simulación teóricos más tarde mostraron que mientras que la selección de grupo es posible, en principio, sólo se puede superar la fuerte selección individual bajo condiciones muy estrictas que rara vez pueden aplicarse en naturaleza.[18] Pero tales condiciones pueden ser creadas en el laboratorio. Cuando escarabajos de la harina *(tribolium castaneum)* fueron criados para reducir el tamaño de la población, mediante la aplicación de una fuerte selección de grupo, la evolución efectivamente condujo a poblaciones menores.[19] Sin embargo, los medios mediante los cuales se logró esto incluyeron no sólo las adaptaciones "benignas" de reducción de la fecundidad y de tiempo de desarrollo más amplio que una búsqueda evolutiva humana ingenuamente antropomorfizadora podría haber esperado, sino también un aumento del canibalismo.[20]

bilidad extra en otras). Esto está mucho muy lejos del alcance de la IA contemporánea. Y debido a la explosión combinatoria, que generalmente echa por tierra los intentos de resolver problemas de planificación complicados con métodos de fuerza bruta (como vimos en el capítulo 1), las deficiencias de los algoritmos conocidos no pueden superarse de manera realista simplemente vertiendo sobre ellos más poder computacional.[21] Sin embargo, una vez que los procesos de planificación o de búsqueda se vuelvan lo suficientemente potentes, también se volverán potencialmente peligrosos.

En lugar de permitir un comportamiento intencional similar al de un agente que surja de manera espontánea y al azar de la aplicación de poderosos procesos de búsqueda (incluyendo los procesos que busquen los planes de trabajo internos y los procesos que busquen directamente soluciones que reúnan algún criterio especificado por el usuario), puede que sea mejor crear agentes a propósito. Dotar a una superin- teligencia con una estructura explícita similar a la de un agente puede ser una manera de aumentar la previsibilidad y la transparencia. Un sistema bien diseñado, construido de tal manera que hubiera una separación limpia entre sus valores y sus creencias, nos permitiría predecir algo acerca de los resultados que tendería a producir. Incluso si no pudiéramos prever exactamente qué creencias adquiriría el sistema o en qué situaciones se encontraría, habría un lugar conocido donde podríamos inspeccionar sus valores finales y,

por lo tanto, los criterios que utilizaría para seleccionar sus futuras acciones y para evaluar cualquier plan potencial.

Comparación

Puede ser útil resumir las características de las diferentes castas de sistema que hemos discutido (tabla 11).

- Métodos de encajamiento inaplicables
- La mayoría de los otros métodos de control de capacidad también son inaplicables (excepto, posiblemente, la integración social o la captura antrópica)

- Domesticidad mayormente inaplicable

- Gran necesidad de que la IA comprenda los verdaderos intereses e intenciones humanas
- Necesidad de acertar al primer intento (aunque esto, en gran medida, aunque posiblemente menor; es cierto para todas las castas)

- Potencialmente una fuente de gran poder para el patrocinador; incluyendo la posibilidad de una ventaja estratégica decisiva

- Una vez activado, no es vulnerable al secuestro por parte del operador, y puede ser diseñado con algún tipo de protección contra el uso insensato

- Se puede utilizar para poner en práctica el “velo de ignorancia" respecto de los resultados (cf. capítulo 13)

- Los métodos de encajamiento pueden ser aplicables, dependiendo de su implementación
- Es probable que los procesos de búsqueda de gran alcance participen en el desarrollo y operatividad de una superinteligencia artificial

- La búsqueda de gran alcance para encontrar una solución que cumpla algún criterio formal puede producir soluciones que cumplan con el criterio de una forma involuntaria y peligrosa

- La búsqueda de gran alcance podría implicar procesos de búsqueda y planificación secundarios internos que podrían encontrar formas peligrosas de ejecutar el proceso de búsqueda principal

Se necesita más investigación para determinar qué tipo de sistema sería más seguro. La respuesta puede depender de las condiciones en que se despliegue la IA. La casta oráculo es obviamente atractiva desde el punto de vista de la seguridad, ya que permitiría que se aplicaran tanto los métodos de control de la capacidad como los métodos de selección de la motivación. Podría, por lo tanto, parecer que ésta simplemente dominaría a la casta soberana, que sólo admitiría los métodos de selección de motivación (excepto en escenarios en los que se cree que el mundo pudiera contener otras superinteligencias poderosas, en cuyo caso podrían aplicarse la integración social o la captura antrópica). Sin embargo, un oráculo podría dar una gran cantidad de poder a su operador, que podría estar dañado o podría aplicar el poder imprudentemente, mientras que un soberano ofrecería cierta protección contra estos peligros. La clasificación por seguridad, por lo tanto, no se determina tan fácilmente.

Un genio puede ser visto como una solución de compromiso entre un oráculo y un soberano —pero no necesariamente una buena solución. En muchos sentidos, implicaría compartir las desventajas de ambos. La aparente seguridad de una IA-herramienta, por su parte, puede ser ilusoria. Para que las herramientas sean lo suficientemente versátiles como para sustituir a agentes superinteligentes, es posible que necesiten desplegar procesos de búsqueda y planificación interna extremadamente poderosos. Comportamientos similares a los de un agente pueden surgir de tales procesos como consecuencias no planeadas. En ese caso, sería mejor diseñar desde el principio el sistema para que fuera un agente, pues así los programadores podrían ver más fácilmente qué criterios acabarían determinando los efectos del sistema.

CAPÍTULO 11

Escenarios multipolares

H

emos visto (en particular en el capítulo 8) la forma amenazadora que podría tomar un resultado unipolar, en el que una sola superinteligencia obtuviera una ventaja estratégica decisiva y la utilizara para establecer una Unidad. En este capítulo, examinamos lo que sucedería en un resultado multi- polar, en una sociedad de post-transición con múltiples agencias superinteligen- tes compitiendo. Nuestro interés por esta clase de escenarios es doble. En primer lugar, como se alude en el capítulo 9, podría pensarse que la integración social ofrece una solución al problema del control. Ya señalamos algunas limitaciones de ese enfoque, y en este capítulo se pinta un cuadro más completo. En segundo lugar, incluso sin que nadie cree una condición multipolar como manera de manejar el problema del control, ese resultado podría ocurrir de todos modos. Entonces, ¿cómo sería tal resultado? La competitiva sociedad resultante no sería necesariamente atractiva, ni de larga duración.

En escenarios de Unidad, lo que sucede después de la transición depende casi por completo de los valores de la Unidad. Así, el resultado podría ser muy bueno o muy malo, dependiendo de cuáles fueran esos valores. Qué valores sean esos depende, a su vez, de si el problema de control quedó resuelto —y el grado en quedó resuelto— en los objetivos del proyecto que creó la Unidad.

Si uno está interesado en que se produzcan escenarios de Unidad, en ese caso sólo existen tres fuentes de información: información sobre asuntos que no pueden estar afectados por las acciones de la Unidad (como las leyes de la física); información sobre los valores instrumentales convergentes; e información que permita predecir o especular acerca de los valores finales que la Unidad tendrá.

En escenarios multipolares, entran en juego un conjunto adicional de limitaciones, las limitaciones que tienen que ver con cómo los agentes interactúan. Las dinámicas sociales que emergen de estas interacciones pueden ser estudiadas utilizando técnicas de teoría de juegos, economía, y teoría de la evolución. Algunos elementos de ciencia política y sociología también son relevantes en la medida en que puedan destilarse y abstraerse algunas de las características más contingentes de la experiencia humana. Aunque sería poco realista esperar que estas limitaciones nos dieran una idea precisa del mundo posterior a la transición, pueden ayudar a identificar algunas posibilidades sobresalientes y poner en cuestión algunas suposiciones infundadas.

Comenzaremos por explorar un escenario económico caracterizado por una regulación a pequeña escala, una fuerte protección de los derechos de propiedad, y una introducción de mentes digitales de bajo coste moderadamente rápida.[1] Este tipo de modelo está estrechamente asociado al economista estadounidense Robin Hanson, quien ha realizado una labor pionera sobre el tema. Más adelante en este capítulo, contemplaremos algunas consideraciones evolutivas y examinaremos las perspectivas de un mundo post-transición inicialmente multipolar que posteriormente se fusionaría en una Unidad.

De caballos y hombres

La inteligencia artificial general podría servir como un sustituto de la inteligencia humana. No sólo podrían las mentes digitales realizar el trabajo intelectual que ahora desempeñan los seres humanos, sino que, una vez equipada con buenos mecanismos o cuerpos robóticos, las máquinas también podrán sustituir el trabajo físico humano. Supongamos que los trabajadores máquina —que pueden ser rápidamente reproducidos— se vuelven más baratos y más capaces que los trabajadores humanos en prácticamente todos los puestos de trabajo. ¿Qué sucedería entonces?

Los salarios y el desempleo

Con mano de obra barata reproducible, los salarios del mercado caen. El único lugar donde los humanos seguirían siendo competitivos sería allí dónde los clientes tengan una preferencia básica por el trabajo realizado por seres humanos. Hoy en día, los productos que han sido elaborados a mano o producidos por indígenas, a veces, se venden a un precio más alto. Los futuros consumidores podrían igualmente preferir los bienes humanos a medida, y a los atletas humanos, a los artistas humanos, a los amantes humanos y a los líderes humanos frente a sus homólogos artificiales funcionalmente indistinguibles o superiores. No está claro, sin embargo, cómo de generalizadas serían tales preferencias. Si las alternativas hechas a máquina fueran suficientemente superiores, quizá tendrían un precio más alto.

Un parámetro que puede ser relevante para la elección del consumidor es la vida interior del trabajador que presta un servicio o producto. A la audiencia de un concierto, por ejemplo, podría gustarle saber que el artista está experimentando conscientemente la música y el lugar de la celebración. En ausencia de experiencia fenoménica, el músico podría ser considerado como una mera máquina de discos de alta resolución, aunque fuera capaz de crear la apariencia tridimensional de un artista interactuando de forma natural con la multitud. Las máquinas podrían entonces ser diseñadas para copiar los estados mentales que estarían presentes en un ser humano que realizara la misma tarea. Sin embargo, incluso con la replicación perfecta de experiencias subjetivas, algunas personas pueden simplemente preferir el trabajo orgánico. Tales preferencias también podrían tener raíces ideológicas o religiosas. Al igual que muchos musulmanes y judíos rechazan la comida preparada de maneras que clasifican como *haram* o *treif,* por lo mismo podría haber grupos en el futuro que evitaran los productos cuya fabricación hubiera implicado el uso no autorizado de inteligencia artificial.

¿Qué importa esto? En la medida en que el trabajo de las máquinas sea barato, éste puede llegar a sustituir al trabajo humano, y los trabajos humanos podrían desaparecer. Los temores sobre la automatización y la pérdida del empleo, por supuesto, no son nuevos. La preocupación por el desempleo tecnológico ha surgido periódicamente, por lo menos desde la Revolución Industrial; y un buen número de profesiones han seguido el camino de los tejedores ingleses y los artesanos textiles que a principios del siglo XIX se unieron bajo la bandera de la folclórica “General Ludd” para luchar contra la introducción de telares mecánicos. Sin embargo, a pesar de que la maquinaria y la tecnología han sustituido a muchos tipos de trabajo humano, la tecnología física ha sido en general un complemento al trabajo humano. Los salarios humanos promedio de todo el mundo han tenido a largo plazo una tendencia alcista, en gran

parte debido a estas complementariedades. Sin embargo, lo que comienza como un complemento al trabajo puede posteriormente convertirse en un sustituto de la mano de obra. Los caballos fueron inicialmente complementados por carros y arados, lo que aumentó considerablemente la productividad del caballo. Más tarde, los caballos fueron sustituidos por automóviles y tractores. Estas innovaciones posteriores redujeron la demanda de trabajo equina y condujeron a un colapso de su población. ¿Podría un destino similar sobrevenir sobre la especie humana?

El paralelismo con la historia del caballo se puede acentuar aún más si nos preguntamos por qué todavía hay caballos en nuestro mundo. Una de las razones es que todavía hay nichos en los que los caballos tienen ventajas funcionales; por ejemplo, el trabajo policial. Pero la razón principal es que los humanos tienen preferencias particulares por los servicios que los caballos pueden ofrecer, incluyendo paseos a caballo de recreo y la competición. Estas preferencias se pueden comparar con las preferencias que algunos humanos podrían tener en este hipotético futuro, de que determinados bienes y servicios estuvieran hechos por manos humanas. Aunque sugerente, esta analogía es, sin embargo, inexacta, ya que todavía no hay sustituto funcional completo para los caballos. Si hubiera dispositivos mecánicos de bajo coste que corrieran a través del heno y tuvieran exactamente la misma forma de sentir, oler, y comportarse como los caballos biológicos —tal vez incluso con las mismas experiencias conscientes— entonces la demanda de caballos biológicos, probablemente se reduciría aún más.

Con una reducción suficiente en la demanda de mano de obra humana, los salarios caerían por debajo del nivel de subsistencia humana. La desventaja potencial de los trabajadores humanos sería, por tanto, extrema: no sólo la reducción salarial, los descensos de categoría, o la necesidad de reconversión, sino el hambre y la muerte. Cuando los caballos se volvieron obsoletos como fuente de energía móvil, muchos fueron vendidos a empacadores de carne para ser transformados en alimentos para perros, harina de hueso, cuero y pegamento. Estos animales no tenían un empleo alternativo a través del cual ganarse su sustento. En los Estados Unidos, había alrededor de 26 millones de caballos en 1915. A principios de la década de 1950, quedaban dos millones.[2]

El capital y el bienestar

Una de las diferencias entre los humanos y los caballos es que los seres humanos poseen capital. Un hecho empírico estilizado es que el

porcentaje total de los factores de capital se mantiene estable desde hace mucho tiempo en aproximadamente el 30% (aunque con importantes fluctuaciones a corto plazo).[3] Esto significa que el 30% de la renta mundial total se recibió en concepto de alquiler por los dueños de capital, el 70% restante se recibe como salario de los trabajadores. Si clasificamos a la IA como capital, entonces, con la invención de una inteligencia artificial que pudiera sustituir completamente al trabajo humano, los salarios caerían hasta el coste marginal de tales sustitutos artificiales, que —bajo el supuesto de que las máquinas fueran muy eficientes— sería muy bajo, muy por debajo de los ingresos del nivel de subsistencia humano. La cuota de ingresos recibidos por la mano de obra iría entonces disminuyendo hasta prácticamente desaparecer. Pero esto implica que la participación del factor del capital se convertiría en casi el 100% del producto mundial total. Puesto que el PIB mundial se dispararía tras una explosión de inteligencia (a causa de las enormes cantidades de nuevas máquinas sustitutivas de la mano de obra, pero también debido a los avances tecnológicos logrados por la superinteligencia, y, más tarde, por la adquisición de grandes cantidades de nuevas tierras a través de la colonización espacial), se deduce que los ingresos totales de capital aumentarían enormemente. Si los seres humanos siguieran siendo los dueños de este capital, el ingreso total recibido por la población humana crecería astronómicamente, a pesar de que en este escenario los seres humanos ya no recibirían ningún ingreso salarial.

Por lo tanto, la especie humana en su conjunto podría llegar a ser más rica que cualquier sueño provocado por la avaricia. ¿Cómo se distribuiría este ingreso? En una primera aproximación, las rentas de capital serían proporcionales a la cantidad de capital en propiedad. Teniendo en cuenta el efecto de amplificación astronómica, incluso un pequeño pedacito de la riqueza anterior a la transición se convertiría en una inmensa fortuna después de la transición. Sin embargo, en el mundo contemporáneo, muchas personas no tienen riqueza. Esto incluye no sólo a las personas que viven en la pobreza, sino también a algunas personas que ganan un buen ingreso o que tienen alto capital humano, pero tienen valor neto negativo. Por ejemplo, en las opulentas Dinamarca y Suecia, el 30% de la población refiere poca riqueza —a menudo, gente joven de clase media, con pocos activos tangibles y con una deuda por su tarjeta de crédito o por sus préstamos universitarios.[4] Incluso si los ahorros produjeran un altísimo interés, se necesitaría una riqueza seminal, algún capital inicial, a fin de que la acumulación pudiera empezar.[5]

Sin embargo, incluso las personas que no tuvieran riqueza privada en el inicio de la transición podrían llegar a ser extremadamente ricas. Los que tuvieran un plan de pensiones, por ejemplo, ya sea público o

privado, deberían estar en una buena posición, siempre que el programa estuviera al menos parcialmente financiado.[6] Los proletarios también podrían hacerse ricos a través de la filantropía de los que ven que su patrimonio neto se dispara: debido al tamaño astronómico de la bonanza, incluso una fracción muy pequeña donada como limosna se convertiría en una gran suma en términos absolutos.

También es posible que las riquezas aún pudieran conseguirse a través del trabajo, incluso en una etapa posterior a la transición cuando las máquinas fueran funcionalmente superiores a los seres humanos en todos los ámbitos (incluso más baratas que el trabajo de nivel de subsistencia humano). Como se señaló anteriormente, esto podría suceder si hubiera nichos en los que se prefiriera el trabajo humano por razones estéticas, ideológicas, éticas, religiosas o razones no pragmáticas de otro tipo. En un escenario en el que la riqueza de los titulares de capital humano aumentara dramáticamente, la demanda de este tipo de trabajo podría aumentar en correspondencia. Nuevos trillonarios o cuatrillonarios podrían permitirse el lujo de pagar una prima considerable por tener algunos de sus bienes y servicios suministrados por mano de obra de "comercio justo" orgánico. La historia de los caballos de nuevo ofrece un paralelismo. Después de caer a dos millones en la década de 1950, la población de caballos de Estados Unidos experimentó una importante recuperación: un censo reciente pone el número en poco menos de diez millones de cabezas.[7] La subida no se debe a nuevas necesidades funcionales de caballos en agricultura o transporte; más bien, se debe a que el crecimiento económico ha permitido a más estadounidenses disfrutar del lujo de la recreación ecuestre.

Otra diferencia relevante entre humanos y caballos, aparte de la propiedad de capital, es que los seres humanos son capaces de movilización política. Un gobierno humano podría usar el poder fiscal del Estado para redistribuir las ganancias privadas, o para aumentar los ingresos mediante la venta de activos estatales valiosos, como la tierra pública, y utilizar las ganancias para las pensiones de sus constituyentes. Una vez más, debido al explosivo crecimiento económico durante e inmediatamente después de la transición, habría mucha más riqueza restante, por lo que sería relativamente fácil abastecer a todos los ciudadanos desempleados. Debería ser posible incluso para un solo país proporcionar a todos los seres humanos de todo el mundo un salario digno generoso sin mayor coste proporcional que lo que muchos países gastan actualmente en ayuda extranjera.

El principio malthusiano desde una perspectiva histórica

Hasta ahora hemos asumido una población humana constante. Esto puede ser una suposición razonable para escalas cortas de tiempo, ya que la biología limita el ratio de reproducción humana. En escalas de tiempo más largas, sin embargo, la suposición no es necesariamente razonable.

La población humana ha aumentado mil veces en los últimos 9.000 años.[9] El aumento habría sido mucho más rápido si no fuera por el hecho de que en la mayor parte de la historia y la prehistoria, la población humana se ha topado con los límites de la economía mundial. Una condición más o menos maltusiana prevaleció, en la que la mayoría de las personas reciben ingresos de subsistencia que apenas les permiten sobrevivir y criar una media de dos hijos hasta la madurez.[10] Hubo bajadas temporales y locales: las plagas, las fluctuaciones del clima, o la guerra mermaron de forma intermitente la población y dejaron tierras libres, lo que permitió a los supervivientes mejorar su ingesta nutricional —y que nacieran más hijos, hasta que se repusieron las filas y la situación malthusiana se restableció. Además, gracias a la desigualdad social, un pequeño estrato elitista pudo disfrutar de ingresos consistentemente por encima del nivel de subsistencia (a expensas de reducir hasta cierto punto el tamaño total de la población que podía ser sostenida). Un pensamiento triste y disonante: que en esta situación maltusiana, el estado normal de las cosas durante la mayor parte de nuestra estancia en este planeta, ha sido de sequías, pestes, masacres, y desigualdad —en la estimación común, los peores enemigos del bienestar humano— las cuales pueden haber sido las más humanitarias: sólo ellas han permitido que el nivel medio de bienestar se elevara ocasionalmente un poco por encima del nivel de subsistencia.

Superpuesta a las fluctuaciones locales, la historia muestra una macro-patrón de crecimiento económico inicialmente lento pero que va acelerando, impulsado por la acumulación de innovaciones tecnológicas. La creciente economía mundial trajo consigo un aumento proporcional de población mundial. (De manera más precisa, una población mayor parece haber acelerado el ritmo de crecimiento, tal vez principalmente por aumentar la inteligencia colectiva de la humanidad)[11]. Sólo a partir de la Revolución Industrial, sin embargo, el crecimiento económico se volvió tan rápido que el crecimiento demográfico no pudo mantener el ritmo. Por tanto, los ingresos promedio comenzaron a subir, primero en los países tempranamente industrializados de la Europa occidental, posteriormente en la mayor parte del mundo. Incluso en los países más pobres de hoy, el ingreso promedio supera sustancialmente el nivel de subsistencia, como se refleja en el hecho de que las poblaciones de estos países estén creciendo.

Los países más pobres tienen ahora un crecimiento de población más rápido, ya que aún tienen que completar la "transición demográfica" al régimen de baja fertilidad que se ha apoderado de las sociedades más desarrolladas. Los demógrafos proyectan que la población mundial se elevará a cerca de nueve mil millones para mediados de siglo, y que a partir de entonces podría asentarse o disminuir a medida que los países más pobres se unieran al mundo desarrollado en este régimen de baja fertilidad.[12] Muchos países ricos ya tienen tasas de fertilidad que están por debajo del nivel de reemplazo; en algunos casos, muy por debajo.[13]

Sin embargo, hay razones, si se tiene una visión más amplia y asumimos un estado de tecnología inalterado y de continua prosperidad, para esperar un retorno a la situación histórica y ecológica normal de una población mundial que topa contra los límites de lo que nuestro nicho puede soportar. Si esto parece contradictorio a la luz de la relación negativa entre la riqueza y la fertilidad que actualmente estamos observando a la escala global, debemos recordar que esta era moderna es un breve trozo de la historia y en gran medida una aberración. El comportamiento humano aún no se ha adaptado a las condiciones contemporáneas. No sólo no somos capaces de aprovechar maneras obvias de aumentar nuestra aptitud inclusiva (como convirtiéndonos en donantes de esperma u óvulos) sino que saboteamos activamente nuestra fertilidad mediante el uso de métodos anticonceptivos. En un entorno de adaptabilidad evolutiva, un impulso sexual saludable puede haber sido suficiente para que un acto individual se realizara de manera que maximizara su potencial reproductivo; en el ambiente moderno, sin embargo, habría una enorme ventaja selectiva en tener un deseo directo de ser el padre biológico del mayor número posible de niños. Tal deseo se está seleccionando actualmente, al igual que otros rasgos que aumentan nuestra propensión a reproducirnos. La adaptación cultural, sin embargo, podría adelantarse a la evolución biológica. Algunas comunidades, como las de los hutteritas o los partidarios del movimiento evangélico Quiverfull, tienen culturas de natalidad que fomentan las familias numerosas, y están, por consiguiente, experimentando una rápida expansión.

Crecimiento demográfico e inversión

Si imaginamos las condiciones socioeconómicas actuales mágicamente congeladas en su actual formato, el futuro estaría dominado por los grupos culturales o étnicos que hubieran mantenido altos niveles de fertilidad. Si la mayoría de la gente tuviera preferencias que maximizaran la vida sana en el entorno actual, la población podría

duplicarse fácilmente en cada generación. En ausencia de políticas de control de población —las cuales tendrían que hacerse cada vez más rigurosas y eficaces para contrarrestar la evolución de fuertes preferencias por eludirlas— la población mundial seguiría entonces creciendo exponencialmente hasta que algún obstáculo, como la escasez de tierras o el agotamiento de oportunidades fáciles para innovaciones importantes, hiciera imposible que la economía mantuviera el ritmo: en ese momento, los ingresos medios empezarían a declinar hasta llegar al nivel en el que la aplastante pobreza impidiera a la mayoría de la gente criar más de dos hijos hasta la madurez. Así, el principio de Mal- thus podría reafirmarse, como un esclavista malvado, con lo que nuestra aventura en el mundo de los sueños de la abundancia llegaría a su fin, volviendo de nuevo a la cantera con las cadenas, para reanudar la agotadora lucha por la subsistencia.

Esta perspectiva a más largo plazo podría ser proyectada a un momento inminente por la explosión de inteligencia. Dado que el software puede copiarse, una población de emulaciones o de IAs podría duplicarse rápidamente —en el transcurso de minutos en lugar de décadas o siglos— agotando enseguida todos los hardwares disponibles.

La propiedad privada puede ofrecer una protección parcial contra la aparición de una situación maltusiana universal. Considérese la posibilidad de un modelo sencillo en el que clanes (o comunidades cerradas, o Estados) comenzaran con cantidades variables de propiedad y de forma independiente adoptaran diferentes políticas sobre la reproducción y la inversión. Algunos clanes descuidan su futuro directamente y gastan sus recursos, con lo cual sus empobrecidos miembros se unen al proletariado mundial (o mueren, si no pueden mantenerse a sí mismos a través de su trabajo). Otros clanes invierten parte de sus recursos, pero adoptan una política de reproducción ilimitada: esos clanes crecen hasta que alcanzan una condición maltusiana interna en la que sus miembros son tan pobres que mueren casi al mismo ritmo que se reproducen, momento en el que el crecimiento de la población del clan ralentiza para igualar el crecimiento con sus recursos. Sin embargo, otros clanes restringirían su fertilidad por debajo de la tasa de crecimiento de su capital: esos clanes lentamente podrían incrementar su número, mientras que sus miembros también se harían más ricos per cápita.

Si la riqueza se redistribuyera de los clanes ricos a los miembros de los clanes que se reproducen rápidamente o que se descuidaron rápidamente (cuyos niños, copias, o vástagos, por causas ajenas a sí mismos, llegaron al mundo con capital insuficiente para sobrevivir y prosperar) entonces nos aproximaríamos lentamente a una situación

maltusiana universal. En el caso límite, todos los miembros de todos los clanes recibirían ingresos de subsistencia y todo el mundo sería igualmente pobre.

Si la propiedad no se redistribuyera, los clanes prudentes podrían aferrarse a una cierta cantidad de capital, y es posible que su riqueza creciera en términos absolutos. No está claro, sin embargo, si los seres humanos podrían conseguir las mismas altas tasas de retorno sobre su capital como las inteligencias artificiales podrían conseguir sobre el suyo, porque puede haber sinergias entre el trabajo y el capital de manera que un agente único que pudiera suministrar ambos (por ejemplo, un empresario o inversionista que fuera a la vez hábil y rico) podría alcanzar una tasa de rentabilidad privada de su capital social superior a la tasa de mercado obtenible por agentes que poseen recursos financieros, pero no cognitivos. Los seres humanos, al ser menos hábiles que las inteligencias artificiales, sólo podrían aumentar su capital lentamente —a menos que, por supuesto, el problema de control hubiera sido completamente resuelto, en cuyo caso la tasa humana de retorno sería igual a la de la máquina, ya que un director humano podría encargar a un agente artificial que gestionara sus ahorros, y podría hacerlo sin coste y sin conflictos de interés: pero, de no ser así, en este escenario la fracción de la economía en propiedad de las máquinas se acercaría asintomáticamente al cien por cien.

Un escenario en el que la fracción de la economía propiedad de las máquinas se aproximara asintomáticamente al cien por cien no es necesariamente uno en el que el tamaño de la parte humana disminuyera. Si la economía creciera a un ritmo suficiente, entonces incluso una fracción relativamente decreciente aún podría estar aumentando su tamaño en términos absolutos. Esto puede sonar como una noticia moderadamente buena para la humanidad: en un escenario multipolar en el que los derechos de propiedad estén protegidos —incluso si fuéramos completamente incapaces de resolver el problema del control— el total de la riqueza en propiedad de los seres humanos podría aumentar. Por supuesto, este efecto no se haría cargo del problema del crecimiento demográfico de la población humana haciendo descender el ingreso per cápita a nivel de subsistencia, ni el problema de los seres humanos que se arruinan a sí mismos por descuidar el futuro.

A largo plazo, la economía se volvería cada vez más dominada por los clanes que tuvieran las más altas tasas de ahorro —avaros que son dueños de media ciudad y viven bajo un puente. Sólo en la plenitud de los tiempos, cuando no haya más oportunidades de inversión, comenzarían los avaros máximamente prósperos a retirar sus ahorros.[14] Sin embargo, si no hubiera una protección perfecta para los derechos de propiedad —por ejemplo, si las máquinas más eficientes

en red tuvieran éxito, por las buenas o por las malas, en transferir la riqueza de los seres humanos a sí mismas— entonces los capitalistas humanos podrían tener que gastar su capital mucho antes, antes de que se agotara por tales transferencias (o por los costes que conllevaría la obtención de su riqueza en contra de tales transferencias). Si estos acontecimientos tuvieran lugar de manera digital en lugar de en escalas de tiempo biológicas, entonces los glaciares seres humanos podrían verse expropiados antes de que pudieran decir Jack Robinson.[15]

La vida en una economía algorítmica

La vida para los seres humanos biológicos en un estado malthusiano de post-tran- sición no tiene por qué parecerse a cualquiera de los estados históricos del hombre (como cazadores-recolectores, agricultores o trabajadores de oficina). En cambio, la mayoría de los seres humanos en este escenario podría ser rentistas ociosos que llevaran su vida adelante gracias a sus ahorros.[16] Serían muy pobres, y lo poco que ingresaran provendría de ahorros o de los subsidios estatales. Vivirían en un mundo con una tecnología muy avanzada, incluyendo no sólo las máquinas superinteligentes sino también la medicina anti-envejecimiento, la realidad virtual, y varias tecnologías de mejora y drogas de recreo: sin embargo éstas podrían generalmente ser inaccesibles. Tal vez en lugar de usar la medicina mejorativa, tomarían las drogas para detener su crecimiento y disminuir su metabolismo con el fin de reducir su coste de vida (los de metabolismo rápido serían incapaces de sobrevivir a la gradual disminución del ingreso de subsistencia). A medida que nuestros números de aumento y nuestro ingreso promedio disminuyera aún más, podríamos degenerar en cualquier mínima estructura susceptible de recibir una pensión —quizás cerebros en probetas mínimamente conscientes, oxigenados y nutridos por máquinas, ahorrando poco a poco el dinero suficiente para reproducirse mediante un técnico robot que desarrolle un clon de nosotros mismos.[17]

Más frugalidad podría lograrse por medio de la subida a la nube, ya que un sustrato de computación físicamente optimizado, ideado por una superinteligencia avanzada, sería más eficiente que un cerebro biológico. La migración al ámbito digital podría detenerse, sin embargo, si las emulaciones fueran consideradas no-humanas o no-ciudadanos incapaces de recibir pensiones o de mantener cuentas de ahorro libres de impuestos. En ese caso, un nicho para los humanos biológicos podría permanecer abierto, junto a una población tal vez mucho más grande de emulaciones o inteligencias artificiales.

Hasta ahora nos hemos centrado en el destino de los seres humanos, que pueden ser apoyados por el ahorro, las subvenciones o los ingresos salariales derivados de otros humanos que prefieran contratar a seres humanos. Volvamos ahora nuestra atención a algunas entidades que hasta ahora hemos clasificado como "capital": máquinas que pueden ser propiedad de los seres humanos, que estarían construidas y enfocadas a cumplir con las tareas funcionales que realizan, y que son capaces de sustituir al trabajo humano en una amplia gama de puestos de trabajo. ¿Cuál puede ser la situación de estos caballos de batalla de la nueva economía?

Si estas máquinas fueran meros autómatas, dispositivos simples como una máquina de vapor o el mecanismo de un reloj, entonces no se necesitarían más comentarios: habría una gran cantidad de ese capital en una economía post-transición, pues a nadie le importaría qué le pasara a piezas de equipo insensibles. Sin embargo, si las máquinas tuvieran mentes conscientes —si estuvieran construidas de tal manera que su operación estuviera asociada con tener una conciencia fenoménica (o si por alguna otra razón se les atribuyera estatus moral)— entonces sería importante tener en cuenta el resultado global en términos de cómo afectaría a estas mentes artificiales. El bienestar de las mentes artificiales de trabajo podría incluso llegar a ser el aspecto más importante del resultado, ya que podrían ser numéricamente dominantes.

Esclavitud voluntaria, muerte ocasional

Una pregunta inicial relevante es si estas mentes-máquina trabajadoras son propiedad como capital (esclavos) o son contratadas como jornaleros libres. En una inspección más cercana, sin embargo, es dudoso que algo realmente dependa de esta cuestión. Hay dos razones para ello. En primer lugar, si a un trabajador libre en un estado malthusiano se le pagara un salario de subsistencia, no tendrá ningún ingreso disponible restante después de haber pagado la comida y otras necesidades. Si el trabajador fuera, en cambio, un esclavo, su dueño pagará por su mantenimiento y tampoco tendrá ingresos disponibles. En cualquier caso, el trabajador conseguiría las necesidades básicas y nada más. En segundo lugar, supongamos que el trabajador libre estuviera de alguna manera en condiciones de proporcionar un ingreso por encima del nivel de subsistencia (tal vez debido a una regulación favorable). ¿En qué gastaría el superávit? A los inversores les resultaría más rentable crear trabajadores que fueran "esclavos voluntarios" —que estuvieran dispuestos a trabajar por salarios de subsistencia. Los inversores podrían crear trabajadores de ese tipo copiando trabajadores que fueran dóciles. Con la selección

apropiada (y quizás algunas modificaciones de código) los inversores podrían ser capaces de crear trabajadores que no sólo prefirieran trabajar gratis, sino que también optaran por donar a sus propietarios los excedentes que pudieran llegar a recibir. Dar dinero al trabajador sería entonces una manera indirecta de dar dinero al propietario o empleador, incluso si el trabajador fuera un agente libre con derechos legales.

Tal vez se objetará que sería difícil diseñar una máquina que quisiera trabajar gratis en cualquier desempeño que se le asignara, o que quisiera donar su salario a su dueño. Podría imaginarse que las emulaciones, en particular, tendrían deseos más típicamente humanos. Pero téngase en cuenta que incluso si el problema de control original fuera difícil, estamos aquí considerando una condición *posterior* a la transición, un momento en el que los métodos para la selección de la motivación probablemente se habrían perfeccionado. En el caso de las emulaciones, se podría llegar muy lejos con sólo *seleccionar* de la gama ya existente de caracteres humanos; y hemos descrito varios otros métodos de selección de motivación. El problema del control puede también, en cierto modo, simplificarse por la suposición actual de que la nueva inteligencia artificial entraría en una matriz socioeconómica estable que ya estaría compuesta de otros agentes superinteligentes respetuosos de la legalidad.

Consideremos, entonces, la difícil situación de la máquina de clase trabajadora, ya sea actuando como un esclavo o como un agente libre. Nos centraremos primero en las emulaciones, el caso más fácil de imaginar.

Traer un nuevo trabajador biológico humano al mundo requiere entre quince y treinta años, dependiendo de la cantidad de conocimientos y experiencia que se requiera. Durante este tiempo, la nueva persona debe ser alimentada, alojada, criada y educada —a un alto coste. Por el contrario, generar una nueva copia de un trabajador digital es tan fácil como cargar un nuevo programa en la memoria de trabajo. Por tanto, la vida se vuelve barata. Una empresa puede adaptar continuamente su fuerza de trabajo para adaptarse a las demandas creando nuevas copias —y eliminando las copias que ya no fueran necesarias, para liberar de recursos al ordenador. Esto podría llevar a una tasa de mortalidad muy alta entre los trabajadores digitales. Muchos podrían vivir sólo un día subjetivo.

Hay razones distintas a las fluctuaciones en la demanda por las que los empresarios o los propietarios de emulaciones podrían querer "matar" o "eliminar" a sus trabajadores frecuentemente.[18] Si una mente de emulación, como una mente biológica, requiriera períodos de descanso y sueño para funcionar, podría ser más barato borrar una emulación fatigada al final de un día y reemplazarla con el estado

almacenado de una emulación fresca y descansada. Como este procedimiento podría causar amnesia retrógrada por todo lo que habrían aprendido durante ese día, las emulaciones que realizaran tareas que requirieran largos hilos cognitivos, se salvarían de dichos borrados frecuentes. Sería difícil, por ejemplo, escribir un libro si cada mañana cuando nos sentáramos en el escritorio no tuviéramos ningún recuerdo de lo que habíamos hecho antes. Pero otros trabajos podrían realizarse adecuadamente por agentes que se reciclaran con frecuencia: un dependiente o un agente de servicio al cliente, una vez entrenado, sólo necesita recordar nueva información durante veinte minutos.

Puesto que el reciclaje de emulaciones impide la formación de memoria y de habilidad, algunas emulaciones podrían colocarse en una pista especial de aprendizaje en la que se ejecutarían de forma continua, incluyendo el descanso y el sueño, incluso en empleos que no requirieran estrictamente largos subprocesos cognitivos. Por ejemplo, algunos agentes de servicio al cliente pueden funcionar durante muchos años en entornos de aprendizaje optimizados, asistidos por entrenadores y evaluadores de desempeño. Los mejores de estos alumnos podrían utilizarse como sementales, sirviendo como plantillas de las que millones de ejemplares frescos saldrían cada día. Se haría un gran esfuerzo para mejorar el desempeño de dichas plantillas de trabajadores, ya que incluso un pequeño incremento de productividad proporcionaría un gran valor económico cuando se aplicara a millones de copias.

Paralelamente a los esfuerzos para capacitar a los trabajadores-plantilla en determinados puestos de trabajo, también se harían intensos esfuerzos para mejorar la tecnología de emulación subyacente. Los avances aquí serían aún más valiosos que los avances en los trabajadores-plantilla individuales, puesto que las mejoras generales de tecnología podrían aplicarse a todos los trabajadores de emulación (y potencialmente a las emulaciones no-trabajadoras también) y no sólo para los que estuvieran en un determinado trabajo. Se dedicarían enormes recursos a la búsqueda de atajos computacionales que permitieran implementaciones más eficientes de emulaciones existentes, así como al desarrollo de arquitecturas de IA neuromórficas y totalmente sintéticas. Esta investigación sería llevada a cabo en su mayoría por emulaciones ejecutándose en hardware muy rápido. Dependiendo del precio de la energía del ordenador, millones, billones o trillones de emulaciones de las mentes humanas más agudas en investigación (o versiones mejoradas de las mismas) podrían estar trabajando todo el día avanzando hacia la frontera de la inteligencia artificial; y algunas de ellas podrían estar operando a varios órdenes de magnitud más rápido que los cerebros biológicos.[19] Ésta es una

buena razón para pensar que la era de las emulaciones de apariencia humana sería breve —un muy breve interludio de tiempo sideral— y que daría paso rápidamente a una era de inteligencia artificial enormemente superior.

Ya hemos encontrado varias razones por las cuales los empleadores de trabajadores de emulación pueden sacrificar periódicamente sus rebaños: fluctuaciones en la demanda de diferentes tipos de trabajadores, el ahorro en costes por no tener que emular el descanso y el tiempo de sueño, y la introducción de plantillas nuevas y mejoradas. Los problemas de seguridad pueden suministrar otra razón. Para evitar que los trabajadores desarrollen planes y conspiraciones subversivas, las emulaciones en posiciones sensibles podrían ser ejecutadas sólo durante períodos limitados, con reseteos frecuentes a estados-base de almacenamiento previo.[20]

Estos estados-base a los que las emulaciones serían reiniciadas estarían cuidadosamente preparados y examinados. Una típica emulación de corta duración podría despertar en un estado mental descansado que estuviera optimizado para la lealtad y la productividad. Ella recordaría haberse graduado de las primeras de su clase después de muchos años (subjetivos) de intensa formación y selección, y haber luego disfrutado de unas vacaciones restauradoras y una buena noche de sueño, de haber luego escuchado un discurso de motivación entusiasta y de música animada, y que ahora está a punto de llegar finalmente a trabajar y hacer todo lo posible por su empleador. Ella no estaría excesivamente preocupada por los pensamientos sobre su inminente muerte al final de la jornada de trabajo. Las emulaciones con neurosis de muerte u otras obsesiones serían menos productivas y no habrían sido seleccionadas.[21] ¿El trabajo máximamente eficiente sería divertido?

Una variable importante en la evaluación de la conveniencia de una condición hipotética de este tipo es el estado hedónico de la emulación promedio.[22] ¿Una típica emulación trabajadora sufriría o disfrutaría de la experiencia de trabajar duro en sus tareas?

Debemos resistir la tentación de proyectar nuestros propios sentimientos sobre la emulación trabajadora imaginaria. La pregunta no es si *usted* se sentiría feliz si tuviera que trabajar constantemente y no pasar tiempo con sus seres queridos —un destino terrible, la mayoría estaría de acuerdo.

Es moderadamente más relevante considerar la actual experiencia hedónica promedio del ser humano durante sus horas de trabajo. Estudios de todo el mundo que preguntaban a los encuestados cómo de felices se sentían, resultaron en que la mayoría se consideraba a sí mismo “bastante feliz” o “muy feliz” (con un promedio de 3.1 en una

escala de 1 a 4).[23] Estudios en el afecto promedio, que preguntaban a los en- cuestados con qué frecuencia habían experimentado recientemente distintos estados afectivos positivos o negativos, tendían a obtener un resultado similar (resultando en un efecto neto de alrededor de 0,52 en una escala de -1 a 1). Hay un efecto moderadamente positivo del ingreso per cápita de un país sobre el bienestar promedio subjetivo.[24] Sin embargo, es peligroso extrapolar estos resultados al estado hedónico de las emulaciones trabajadoras futuras. Una de las razones que podrían aducirse es que su condición sería muy diferente: por un lado, podrían estar trabajando mucho más duro; por el contrario, podrían estar libres de enfermedades, dolores, hambre, olores nocivos, etc. Sin embargo, tales consideraciones no dan en el blanco. La consideración más importante aquí es que el tono hedónico sería fácil de ajustar a través del equivalente digital de fármacos o neurocirugía. Esto significa que sería un error inferir el estado hedónico de las futuras emulaciones de las condiciones externas de sus vidas, a través de imaginar cómo nosotros mismos y otras personas como nosotros nos sentiríamos en esas circunstancias. El estado hedónico sería una cuestión de elección. En el modelo que estamos considerando actualmente, la elección se haría por los propietarios del capital que buscan maximizar la rentabilidad de su inversión en las emulaciones trabajadoras. En consecuencia, la cuestión de cómo de felices se sentirían las emulaciones se reduce a la cuestión de qué estados hedónicos serían más productivos (en los diferentes puestos de trabajo en que las emulaciones serían empleadas).

Aquí, de nuevo, se podría intentar hacer inferencias a partir de observaciones sobre la felicidad humana. Si fuera el caso, de que en la mayoría de los tiempos, lugares y ocupaciones, la gente es feliz de manera al menos moderada, tendríamos cierta presunción a favor de sostener la misma postura en un escenario post-transición como el que estamos considerando. Para ser claros, el argumento en este caso no sería que las mentes humanas tienen una predisposición hacia la felicidad por lo que probablemente encontrarían satisfacción bajo estas nuevas condiciones; sino más bien que un cierto nivel medio de felicidad habría demostrado ser adaptable para las mentes humanas en el pasado, así que tal vez un nivel similar de la felicidad demostrará ser adaptable para las mentes de apariencia humana en el futuro. Sin embargo, esta formulación también pone de manifiesto la debilidad de la inferencia: a saber, que las disposiciones mentales que eran adaptables para los homínidos cazadores-recolectores que recorrían la sabana africana no serían necesariamente adaptables para emulaciones modificadas que vivieran en realidades virtuales post-transición. Sin duda, podemos *esperar* que las futuras emulaciones

trabajadoras sean tan felices como, o más felices que, lo que fueron típicamente los trabajadores en la historia humana; pero todavía no hemos visto ninguna razón de peso para suponer que esto sería así (en el escenario multipolar laissez-faire actualmente estudiado).

Considérese la posibilidad de que la razón de la felicidad frecuente entre los humanos (en cualquier limitado sentido en que sea frecuente) es que el estado de ánimo alegre cumple una función de señalización en un entorno de adaptabilidad evolutiva. Dar la impresión a los demás miembros del grupo social de estar en condiciones óptimas —en buen estado de salud, en buena relación con los compañeros, y en la espera confiada de que la buena fortuna continúe— pudo haber impulsado la popularidad de un individuo. Un sesgo hacia la alegría podría por lo tanto haber sido seleccionado, resultando en que la neuroquímica humana está sesgada hacia el afecto positivo en comparación con lo que habría sido la máxima eficiencia en función de criterios materialistas más simples. Si este fuera el caso, entonces el futuro de la *joie de vivre* podría depender de que la alegría conservara inalterada su función de señalización social en el mundo post-transición: un problema sobre el que volveremos en breve.

¿Qué pasa si las almas alegres disiparan más energía que las sombrías? Tal vez los alegres son más propensos a saltos creativos y derroches lujosos —comportamientos que los futuros empleadores podrían minusvalorar en la mayoría de sus trabajadores. Tal vez una fijación hosca o ansiosa en simplemente seguir adelante con el trabajo sin cometer errores sería la actitud maximizadora de productividad en la mayoría de líneas de trabajo. La pretensión no es que esto sea así, sino que no sabemos que no sea así. No obstante, debemos considerar las horribles consecuencias de que tal hipótesis pesimista sobre un futuro estado maltusiano resultara ser cierta: no sólo por el coste de oportunidad de haber podido crear algo mejor —que sería enorme— sino también porque ese estado podría ser malo en sí mismo, posiblemente mucho peor que el estado original de Malthus.

Rara vez nos esforzamos hasta el límite. Cuando lo hacemos, a veces es doloroso. Imagínese correr en una cinta con pendiente empinada —el corazón saliéndose del pecho, los músculos doloridos, los pulmones respirando con dificultad. Una mirada al temporizador: su próxima escapada, que también será su muerte, será en 49 años, 3 meses, 20 días, 4 horas, 56 minutos y 12 segundos. Desearía no haber nacido.

Una vez más, la tesis no es que esto sea lo que va a suceder, sino que no sabemos que no lo sea. Uno podría proponer un caso más optimista. Por ejemplo, no hay ninguna razón obvia de que las emulaciones tuvieran que sufrir lesiones corporales y enfermedades: la

eliminación de la miseria física sería una gran mejora sobre el actual estado de cosas. Además, puesto que cosas hechas de realidad virtual pueden ser bastante baratas, las emulaciones pueden trabajar en entornos suntuosos —en espléndidos palacios, en las cimas de una montaña, en las terrazas situadas en un bosque en primavera, o en las playas de una laguna azul— con la iluminación, la temperatura, el paisaje y la decoración perfectos; libres de humos molestos, ruidos, corrientes de aire e insectos zumbando; vestidos con ropa cómoda, con sensación de limpieza y centrados, y bien alimentados. Más significativamente, si —como parece perfectamente posible— el estado mental humano óptimo para la productividad en la mayoría de los puestos de trabajo es el de alegre entusiasmo, entonces la era de la economía de emulación podría ser bastante paradisíaca.

Habría, en todo caso, un gran valor optativo en preparar las cosas de tal manera que alguien o algo pudiera intervenir para reformar las cosas si la trayectoria por defecto virara hacia la distopía. También podría ser conveniente disponer de algún tipo de puerta de escape que permitiera la escapatoria de la muerte y el olvido si la calidad de la vida se hundiera permanentemente por debajo de un nivel en el que la aniquilación se volviera preferible a continuar en la existencia.

¿Subcontratados inconscientes?

A largo plazo, cuando la era de la emulación dejara paso a la era de la inteligencia artificial (o si la inteligencia artificial se obtuviera directamente a través de la IA sin una etapa de emulación de cerebro completo anterior), el dolor y el placer podrían posiblemente desaparecer por completo en un escenario multipolar, ya que un mecanismo de recompensa hedónica podría no ser el sistema de motivación más eficaz para un agente artificial complejo (que, a diferencia de la mente humana, no tiene el lastre de la herencia del wetware animal). Tal vez un sistema de motivación más avanzado se basaría en la representación explícita de una función de utilidad o en alguna otra arquitectura que no tuviera análogos funcionalmente exactos al placer y al dolor.

Un resultado multipolar un poco más radical —uno que podría implicar la eliminación en un futuro de casi todos los valores— es que el proletariado universal ni siquiera fuera consciente. Esta posibilidad es más destacable con respecto a la IA, que podría ser estructurada de manera muy diferente a la inteligencia humana. Pero incluso si la inteligencia artificial se lograra inicialmente a través de la emulación de cerebro completo, resultando en mentes digitales conscientes, las desatadas fuerzas de la competencia en una economía post-transición podrían fácilmente conducir a la aparición de formas cada vez menos

neuromórficas de inteligencia artificial, ya sea porque se crearían IA sintéticas de novo o porque las emulaciones se alejarían cada vez más, a través de modificaciones y mejoras sucesivas, de su forma humana original.

Consideremos un escenario en el que, después de que la tecnología de emulación se hubiera desarrollado, el progreso continuado en neurociencia e informática (acelerado por la presencia de mentes digitales que sirven tanto de investigadores como de sujetos de prueba) permitiera aislar los módulos cognitivos individuales de una emulación y conectarlos a los módulos aislados de otras emulaciones. Un período de formación y ajuste puede ser necesario antes de que diferentes módulos puedan colaborar de manera efectiva; pero los módulos que se ajusten a las normas comunes podrían interactuar más rápido con otros módulos estándar. Esto haría que los módulos estandarizados fueran más productivos, y crearía una presión para que hubiera una mayor estandarización.

Las emulaciones podrían entonces comenzar a externalizar crecientes porciones de su funcionalidad. ¿Por qué aprender aritmética cuando se puede enviar su tarea numérica de razonamiento a Módulos de Gauss, Inc.? ¿Por qué ser elocuente cuando se puede contratar a Conversaciones Coleridge para poner sus pensamientos en palabras? ¿Por qué tomar decisiones sobre su vida personal cuando existen módulos ejecutivos certificados que pueden escanear su sistema de objetivos y gestionar sus recursos para lograr sus metas mejor que si tratara de hacerlo usted mismo? Algunas emulaciones pueden preferir conservar la mayor parte de su funcionalidad y manejar ellos mismos las tareas que pudieran ser hechas de manera más eficiente por otros. Esas emulaciones serían como los aficionados que disfrutan de plantar sus propias verduras o tejer sus propias chaquetas de punto. Tales emulaciones aficionadas serían menos eficientes; y si hay un flujo neto de recursos entre los participantes menos y más eficientes de la economía, los aficionados finalmente saldrían perdiendo.

Los caldos de cultivo de los distintos intelectos humanos se fundirían así en una sopa algorítmica.

Es concebible que la eficiencia óptima se alcanzara mediante la agrupación de las capacidades en agregados que aproximadamente coincidieran con la arquitectura cognitiva de la mente humana. Podría ser el caso, por ejemplo, que un módulo de matemáticas debiera adaptarse a un módulo de idioma, y que ambos debieran adaptarse al módulo ejecutivo, a fin de que los tres trabajaran juntos. La externalización cognitiva sería entonces casi totalmente inviable. Pero en ausencia de cualquier razón de peso para estar seguros de que esto fuera así, debemos aceptar la posibilidad de que las arquitecturas

cognitivas similares a las humanas son óptimas sólo dentro de las limitaciones de la neurología humana (o no lo son en absoluto). Cuando se haga posible la construcción de arquitecturas que no puedan aplicarse también a redes neuro- nales biológicas, un nuevo espacio de diseño se abrirá; y la optimización mundial en este espacio ampliado no tiene por qué parecerse a ningún tipo familiar de mentalidad. Organizaciones cognitivas similares a las humanas carecerían entonces de nicho en una economía o ecosistema competitivo post-transicional.[25]

Puede haber nichos para complejos que fueran o menos complicados (como módulos individuales), o más complicados (como grandes grupos de módulos) o de complejidad similar a las mentes humanas, pero con arquitecturas radicalmente diferentes. ¿Tienen algún valor intrínseco estos complejos? ¿Hay que dar la bienvenida a un mundo en el que tales complejos alienígenas hubieran sustituido a los complejos humanos?

La respuesta puede depender de la naturaleza específica de los complejos alienígenas. El mundo actual tiene muchos niveles de organización. Algunas entidades altamente complejas, como las corporaciones multinacionales y los Estados-nación, contienen a los seres humanos como componentes; sin embargo, por lo general, sólo asignamos a estos complejos de alto nivel un valor instrumental. Las corporaciones y los Estados no tienen (suponemos generalmente) conciencia, más allá de la conciencia de las personas que los constituyen: no pueden sentir dolor o placer fenoménico o experimentar cualquier qualia. Nosotros los valoramos en la medida en que sirven a las necesidades humanas, y cuando dejan de hacerlo nosotros les "matamos" sin escrúpulos. También hay entidades de nivel más bajo, a las que también se les suele negar el estatus moral. No vemos ningún daño en borrar una aplicación de un teléfono inteligente, y no creemos que un neurocirujano maltrata a nadie cuando extirpa un módulo malfuncionante de un cerebro epiléptico. En cuanto a los complejos exóticamente organizados de un nivel similar al del cerebro humano, la mayoría de nosotros tal vez sólo les atribuiríamos significado moral si pensáramos que tienen capacidad de experiencia consciente.[26]

Así podríamos imaginar, como caso extremo, una sociedad tecnológicamente muy avanzada, que contuviera muchas estructuras complejas, algunas de ellas mucho más complejas e inteligentes que todo lo que existe hoy en día en el planeta, una sociedad que, sin embargo, careciera de cualquier tipo de ser que fuera consciente o cuyo bienestar tuviera significado moral. En cierto sentido, esto sería una sociedad deshabitada. Sería una sociedad de milagros económicos y genialidad tecnológica, con nadie que pudiera beneficiarse. Un

Disneyland sin niños.

La evolución no es necesariamente hacia adelante

La palabra "evolución" se utiliza a menudo como sinónimo de "progreso", quizás reflejando una imagen acrítica común de la evolución como fuerza del bien. Una fe fuera de lugar en la inherente beneficencia del proceso evolutivo puede dificultar una evaluación justa sobre la conveniencia de un resultado multipolar en el que el futuro de la vida inteligente estuviera determinado por una dinámica competitiva. Cualquier evaluación debe basarse en alguna opinión (al menos implícitamente) acerca de la distribución de probabilidad de los diferentes fenotipos que resulten ser adaptativos en una sopa de vida digital post-transición. Sería difícil, en el mejor de los casos, extraer una respuesta clara y correcta del inevitable mejunje de incertidumbre que prevalece sobre estos asuntos: más aún si añadimos una capa de lodo panglossiano.

Una fuente posible para la fe en la evolución hacia adelante es la direccionalidad hacia arriba aparentemente exhibida por el proceso evolutivo en el pasado. A partir de replicadores rudimentarios, la evolución produce cada vez más organismos "avanzados", incluyendo criaturas con mente, conciencia, lenguaje, y razón. Más recientemente, los procesos culturales y tecnológicos, que tienen algunas lejanas similitudes con la evolución biológica, han permitido a los seres humanos desarrollarse a un ritmo acelerado. En una escala de tiempo geológica, así como histórica, el panorama parece mostrar una tendencia general hacia el aumento de los niveles de complejidad, el conocimiento, la conciencia y la organización coordinada dirigida a un objetivo: una tendencia que, si no somos muy quisquillosos, podríamos llamar "progreso"[27]

La imagen de la evolución como un proceso que produce de forma fiable efectos benignos es difícil de conciliar con el enorme sufrimiento que vemos en el mundo humano y natural. Los que aprecian los logros de la evolución puede hacerlo más desde una perspectiva estética que de una perspectiva ética. Sin embargo, la pregunta pertinente no es qué tipo de futuro sería fascinante leer en una novela de ciencia ficción o ver representado en un documental de naturaleza, sino en qué tipo de futuro sería bueno vivir: dos cosas muy diferentes.

Por otra parte, no tenemos ninguna razón para pensar que cualquier progreso que se haya producido fuera en ningún sentido inevitable. Mucho podría haber sido fortuito. Esta objeción gana apoyo en el hecho de que un efecto de selección por observación filtra la evidencia que podemos tener sobre el éxito de nuestro propio

desarrollo evolutivo.[28] Supongamos que el 99,9999% de todos los planetas donde la vida surgió se extinguieron antes de desarrollarse hasta el punto en que los observadores inteligentes pudieran comenzar a reflexionar sobre su origen. ¿Qué podríamos esperar observar si ese fuera el caso? Posiblemente, habría que esperar observar algo parecido a lo que de hecho observamos. La hipótesis de que las probabilidades de que haya vida inteligente en evolución en un planeta dado son bajas no predicen lo que deberíamos encontrarnos en un planeta donde la vida se extinguió en una fase temprana; más bien, se puede predecir que nos encontramos en un planeta donde la vida inteligente evolucionó, incluso aunque tales planetas constituyeran una fracción muy pequeña de todos los planetas donde la vida primitiva evolucionó. La larga trayectoria de la vida en la Tierra, por lo tanto, puede ofrecer escaso apoyo a la afirmación de que había una alta probabilidad —mucho menos algo cercano a la inevitabilidad— implicada en el surgimiento de organismos complejos en nuestro planeta.[29]

En tercer lugar, incluso si las condiciones presentes hubieran sido idílicas, e incluso si pudiera demostrarse que surgieron inevitablemente de algún estado primordial genérico, todavía no habría ninguna garantía de que la tendencia meliorista fuera a continuar en un futuro indefinido. Esto es válido incluso si hacemos caso omiso a la posibilidad de un evento de extinción catastrófica e incluso si asumimos que los desarrollos evolutivos continuarán produciendo sistemas de complejidad creciente.

Hemos sugerido anteriormente que los trabajadores de inteligencia artificial seleccionados para la máxima productividad trabajarían muy duro y que no se sabe lo felices que serían. También nos planteamos la posibilidad de que las formas de vida más aptas en una futura sopa de vida digital competitiva podrían incluso no ser conscientes. A falta de una completa pérdida de placer, o de conciencia, podría haber un desgaste de otras cualidades que muchos podrían considerar como indispensables para una buena vida. Los seres humanos valoran la música, el humor, el romance, el arte, el juego, la danza, la conversación, la filosofía, la literatura, la aventura, el descubrimiento, los alimentos y bebidas, la amistad, la crianza de los hijos, el deporte, la naturaleza, la tradición y espiritualidad, y muchas otras cosas. No hay ninguna garantía de que ninguna de éstas siguieran siendo adaptativas. Tal vez lo que maximizara la aptitud fuera nada más que el trabajo monótono, sin parar, de alta intensidad, el trabajo de carácter monótono y repetitivo, desprovisto de escalofrío lúdico, destinado sólo a la mejora de la octava posición decimal de alguna medida de producción económica. Los fenotipos seleccionados tendrían vidas carentes de las cualidades antes mencionadas, y en

función de la propia axiología el resultado podría parecemos aborrecible, sin valor, o simplemente empobrecido, pero en todo caso muy lejos de una utopía que fuera digna de recomendación.

Cabe preguntarse cómo puede ser consistente un cuadro tan sombrío con el hecho de que nosotros en este momento nos entreguemos a la música, el humor, el romance, el arte, etc. Si estas conductas fueran realmente tan "derrochadoras", entonces ¿cómo podrían haber sido toleradas y de hecho promovidas por los procesos evolutivos que dieron forma a nuestra especie? Que el hombre moderno esté en desequilibrio respecto de la evolución no da cuenta de esto; pues nuestros antepasados del Pleistoceno también participaban en la mayoría de estos divertimentos. Muchos de los comportamientos en cuestión ni siquiera son exclusivos del *homo sapiens.* Una muestra vistosa de esto la encontramos en una amplia variedad de contextos, desde la selección sexual en el reino animal a los enfrentamientos por prestigio entre los Estados-nación.[30]

Aunque una explicación evolutiva completa para cada una de estas conductas está más allá del alcance de la presente investigación, se puede hacer notar que algunas de ellas cumplen funciones que pueden no ser muy relevantes en un contexto de la inteligencia artificial. El juego, por ejemplo, que sólo se produce en algunas especies y predominantemente entre los jóvenes, es principalmente una forma de que el animal joven aprenda las habilidades que necesitará en el futuro. Cuando las emulaciones puedan ser creadas como adultos, ya en posesión de un repertorio maduro de habilidades, o cuando los conocimientos y técnicas adquiridos por una IA se puedan trasladar directamente a otra IA, la necesidad del comportamiento juguetón podría estar menos generalizada.

Muchos de los otros ejemplos de conductas humanas podrían haber evolucionado como señales difíciles de fingir de cualidades que son difíciles de observar directamente, como la resistencia corporal o mental, el estatus social, la calidad de los aliados, la capacidad y la voluntad de prevalecer en una pelea, o la posesión de recursos. La cola del pavo real es el caso clásico: sólo los pavos reales aptos pueden permitirse hacer brotar un plumaje verdaderamente extravagante, y las pavas han evolucionado para encontrarlo atractivo. No menos que los rasgos morfológicos, los rasgos de comportamiento también pueden ser señal de aptitud genética o de otros atributos socialmente relevantes.[31]

Teniendo en cuenta que este comportamiento vistoso es tan común entre los seres humanos como en otras especies, podría cuestionarse si no sería también parte del repertorio de formas de vida tecnológicamente más avanzadas. Incluso si no hubiera ningún uso estrictamente instrumental para la alegría o la musicalidad, o incluso

para la conciencia, en la lógica del procesamiento inteligente de información del futuro, ¿no podrían estos rasgos, sin embargo, conferir alguna ventaja evolutiva a sus poseedores en virtud de ser señales fiables de otras cualidades adaptativas?

Si bien la posibilidad de una armonía preestablecida entre lo que es valioso para nosotros y lo que sería adaptativo en un futuro ecológico digital es difícil de descartar, hay razones para el escepticismo. Consideremos, en primer lugar, que muchas de las costosas muestras que encontramos en la naturaleza están vinculadas a la selección sexual.[32] La reproducción entre formas de vida tecnológicamente maduras, por el contrario, puede ser predominantemente o exclusivamente asexual.

En segundo lugar, los agentes tecnológicamente avanzados pueden tener a su disposición nuevos medios de comunicar información sobre sí mismos de manera fiable, medios que no se basen en exhibiciones costosas. Incluso hoy en día, cuando los prestamistas profesionales evalúan la solvencia tienden a confiar más en pruebas documentales, tales como certificados de propiedad y estados de cuenta bancarios, que en las exhibiciones costosas, tales como trajes de diseñador y relojes Rolex. En el futuro, podría ser posible emplear firmas de auditoría que verificaran mediante el examen detallado de un historial de comportamiento, de pruebas en entornos simulados, o de la inspección directa del código fuente, que un agente cliente poseyera un atributo reivindicado. La señalización de las propias cualidades aceptando dicha auditoría podría ser más eficiente que la señalización a través de comportamientos extravagantes. Dicha señal mediada profesionalmente seguiría siendo costosa de *fingir* —siendo ésta la característica esencial que hace a la señal confiable— pero podría ser mucho más barata de transmitir cuando fuera verdad de lo que costaría comunicar una señal ostentosa.

En tercer lugar, no todas las posibles muestras costosas son intrínsecamente valiosas o socialmente deseables. Muchas son simplemente un desperdicio. Las ceremonias potlatch de los Kwakiutl, una forma de confrontación de estatus entre jefes rivales, implicaban la destrucción pública de grandes cantidades de riqueza acumulada.[33] Rascacielos rompedores de récord, mega-yates y cohetes lunares pueden ser considerados análogos contemporáneos. Si bien actividades como la música y el humor plausiblemente podrían ser reclamados para mejorar la calidad intrínseca de la vida humana, es dudoso que un reclamo similar fuera a hacerse respecto de la costosa búsqueda de accesorios de moda y otros símbolos de estatus consumistas. Peor aún, la exhibición costosa puede ser francamente perjudicial, como en el postureo de machote que conduce a la violencia pandillera o a la bravata militar. Incluso si las futuras formas de vida inteligente usaran

una costosa señalización, es una cuestión abierta si la señal sería valiosa —si sería como la melodía entusiasta de un ruiseñor o en su lugar como el croar monosilábico del sapo (o el incesante ladrido de un perro rabioso).

¿Formación post-transición de una Unidad?

Incluso si el resultado inmediato de la transición a la inteligencia artificial fuera mul- tipolar, todavía quedaría la posibilidad de que una Unidad se desarrollara posteriormente. Tal desarrollo continuaría una aparente tendencia a largo plazo hacia mayores integraciones políticas, llevándola a su conclusión natural.[34] ¿Cómo podría ocurrir esto?

Una segunda transición

El modo en que un resultado inicialmente multipolar podría converger en una Unidad post-transición sería, si tuviera lugar, una segunda transición tecnológica después de la transición inicial lo suficientemente grande y abrupta como para darle una ventaja estratégica decisiva a uno de los poderes restantes: un poder que podría entonces aprovechar la oportunidad de establecer una Unidad. Tal segunda transición hipotética podría ser ocasionada por un gran avance hacia un nivel más alto de su- perinteligencia. Por ejemplo, si la primera ola de superinteligencia artificial estuviera basada en la emulación, a continuación, podría tener lugar una segunda oleada cuando las emulaciones investigadoras tuvieran éxito desarrollando una inteligencia artificial auto-mejorativa efectiva.[35] (Alternativamente, una segunda transición podría ser desencadenada por una avance en nanotecnología o en alguna otra tecnología militar o de propósito general aún no imaginada).

El ritmo de desarrollo después de la transición inicial sería extremadamente rápido. Incluso una pequeña distancia entre la primera potencia y su competidor más cercano podría plausiblemente resultar en una ventaja estratégica decisiva para la primera potencia durante una segunda transición. Supongamos, por ejemplo, que dos proyectos entraran en la primera transición a pocos días de diferencia, y que el despegue fuera lo suficientemente lento como para que esta brecha no diera al proyecto adelantado una ventaja estratégica decisiva en ningún momento durante el despegue. Los dos proyectos emergerían como poderes superinteligentes, aunque uno de ellos permanecería unos días por delante del otro. Pero los acontecimientos

en investigación se estarían produciendo a escalas de tiempo típicas de la superinteli- gencia artificial —quizás miles o millones de veces más rápido que la investigación llevada a cabo en una escala temporal humana biológica. Por lo tanto, el desarrollo de la tecnología de segunda transición podría completarse en días, horas o minutos. A pesar de que la ventaja del adelantado fuera de tan sólo unos días, un gran avance podría, no obstante, catapultarla a una ventaja estratégica decisiva. Téngase en cuenta, sin embargo, que si la difusión tecnológica (mediante el espionaje u otros canales) acelerara tanto como el desarrollo tecnológico, este efecto podría entonces verse anulado. Lo que seguiría siendo relevante sería la subitaneidad de la segunda transición, es decir, la velocidad a la que se desarrollara en relación con la velocidad general de los acontecimientos en el período posterior a la primera transición. (En este sentido, cuanto más rápido sucedan las cosas después de la primera transición, menos abrupta tenderá a ser la segunda transición).

También podríamos especular que, en realidad, sería más probable utilizar una ventaja estratégica decisiva para establecer una Unidad si surgiera durante una segunda (o posterior) transición. Después de la primera transición, los que tomaran las decisiones podrían, o bien ser superinteligentes, o bien tener acceso a asesoramiento superinteligente, lo que aclararía las implicaciones de las opciones estratégicas disponibles. Por otra parte, la situación después de la primera transición podría ser una en la cual un movimiento preventivo contra posibles competidores sería la decisión menos peligrosa para el agresor. Si las mentes de toma de decisiones tras la primera transición fueran digitales, podrían ser copiadas y, por lo tanto, serían menos vulnerables a un contraataque. Incluso si un defensor tuviera la capacidad de matar a nueve décimas partes de la población del agresor en un ataque de represalia, esto no ofrecería apenas disuasión si el compañero fallecido pudiera ser inmediatamente resucitado desde copias de seguridad redundantes. La devastación de la infraestructura (que pudiera ser reconstruida) también podría ser tolerable para mentes digitales con esperanzas de vida efectivamente ilimitadas, que podrían estar planeando maximizar sus recursos e influencia en una escala de tiempo cosmológica.

Superorganismos y economías de escala

El tamaño de agregados humanos coordinados, como las empresas o naciones, están influenciados por diversos parámetros —tecnológicos, militares, financieros y culturales— que pueden variar de una época histórica a otra. Una revolución de inteligencia artificial implicaría cambios profundos en muchos de estos parámetros. Quizás estos

cambios facilitarían la aparición de una Unidad. Aunque no podemos, sin considerar detalladamente en qué consisten estos cambios potenciales, excluir la posibilidad opuesta —que los cambios facilitarían la fragmentación en lugar de la unificación— podemos, sin embargo, observar que el aumento de la variabilidad o de la incertidumbre al que nos enfrentamos aquí puede ser en sí mismo un motivo para dar mayor credibilidad a la posible aparición de una Unidad de lo que hubiéramos pensado en un principio. Una revolución en inteligencia artificial podría, por así decirlo, hacer que las cosas sucedieran —podría volver a barajar el mazo para posibilitar realineamientos geopolíticos que, de otra manera, no parecía que estuvieran en las cartas.

Un análisis exhaustivo de todos los factores que pueden influir la escala de integración política nos llevaría mucho más allá del alcance de este libro: una revisión de la literatura relevante de ciencias políticas y economía podría fácilmente llenar un volumen completo. Debemos limitarnos a hacer una breve alusión a un par de factores, aspectos de la digitalización de los agentes que pueden hacer que sea más fácil centralizar el control.

Carl Shulman ha argumentado que en una población de emulaciones, las presiones de selección favorecerían la aparición de "superorganismos", grupos de emulaciones listas para sacrificarse por el bien de su clan.[36] Los superorganismos se librarían de los problemas de agencia que acosan a las organizaciones cuyos miembros persiguen su propio interés. Al igual que las células de nuestro cuerpo, o los animales individuales en una colonia de insectos eusociales, las emulaciones que fueran totalmente altruistas hacia sus hermanos-copia cooperarían entre sí, incluso en ausencia de planes de incentivos elaborados.

Los superorganismos tendrían una ventaja particularmente fuerte si la eliminación no consensuada (o la suspensión indefinida) de emulaciones individuales fuera rechazada. Las empresas o países que emplearan emulaciones que insistieran en mantener el instinto de conservación se cargarían con el compromiso interminable de pagar el mantenimiento de los trabajadores obsoletos o redundantes. Por el contrario, las organizaciones cuyas emulaciones voluntariamente se borraran a sí mismas cuando sus servicios ya no fueran necesarios podrían adaptarse más fácilmente a las fluctuaciones de la demanda; y podrían experimentar libremente, proliferando variaciones en sus trabajadores y quedándose sólo con los más productivos.

Si la eliminación involuntaria *no* fuera anulada, entonces la ventaja comparativa de las emulaciones eusociales se reduciría, aunque quizás no se eliminaría. Empresarios de cooperativas auto-sacrificadoras aún podrían obtener mejoras en la eficiencia respecto

de los problemas de agencia reducida en toda la organización, incluyendo el ahorrarle la molestia de vencer cualquier resistencia que las emulaciones pudieran poner en contra de su propia eliminación. En general, los aumentos de productividad por tener trabajadores dispuestos a sacrificar sus vidas individuales por el bien común son un caso especial de los beneficios que una organización puede obtener de tener miembros que se dedican fanáticamente a ella. Dichos miembros no sólo saltarían a la tumba por la organización, y trabajarían largas horas por poco dinero: también rechazarían las políticas de oficina y tratarían constantemente de actuar en lo que creyeran ser el mejor interés de la organización, lo que reduciría la necesidad de supervisiones y restricciones burocráticas.

Si la única manera de lograr tanta dedicación fuera restringiendo la afiliación a los hermanos-copia (de modo que todas las emulaciones de un superorganismo particular estuvieran sacadas de la misma plantilla), entonces los superorganismos estarían en desventaja por sólo ser capaces de poner en juego una gama de habilidades más reducidas que las de las organizaciones rivales, una desventaja que podría o no ser lo suficientemente grande como para compensar las ventajas de evitar problemas de agencia interna.[37] Esta desventaja se vería enormemente aliviada si un superorganis- mo pudiera al menos contener miembros con diferente preparación. Incluso si todos sus miembros se derivaran de una única plantilla, su fuerza de trabajo podría aun así contribuir con diversas habilidades. A partir de una plantilla que emulara talento y erudición, se podrían separar linajes en diferentes programas de formación, una copia aprendería contabilidad, otra ingeniería eléctrica, y así sucesivamente. Esto produciría una asociación con diversas habilidades aunque no con diversos talentos. (La diversidad máxima podría requerir que se utilizara más de una plantilla).

La propiedad esencial de un superorganismo no es que se componga de copias de un solo progenitor, sino que todos los agentes individuales dentro de ella están totalmente comprometidos con un objetivo común. La capacidad de crear un superor- ganismo, por lo tanto, puede ser vista como algo que requiere una solución parcial al problema de control. Mientras que una solución general al problema de control permitiría a alguien crear un agente con cualquier objetivo final arbitrario, la solución parcial necesaria para la creación de un superorganismo simplemente requeriría la capacidad de modelar múltiples agentes con un mismo objetivo final (para algunos objetivos finales no triviales pero necesariamente arbitrarios).[38]

La consideración principal presentada en este apartado, por lo tanto, no está en realidad limitada a grupos de emulación monoclonales, sino que puede afirmarse más en general, de manera

que deje claro que se aplica a una amplia gama de escenarios multipolares de inteligencia artificial. Sucede que ciertos tipos de avances en las técnicas de selección de la motivación, que pueden ser factibles cuando los actores sean digitales, pueden ayudar a superar algunas de las ineficiencias que actualmente obstaculizan a las grandes organizaciones humanas y que contrapesan a las economías de escala. Con estos límites superados, las organizaciones —ya sean empresas, naciones, u otras entidades económicas o políticas— podrían aumentar de tamaño. Éste es un factor que podría facilitar la aparición de una Unidad post-transición.

Una de las áreas en la que los superorganismos (u otros agentes digitales con motivaciones parcialmente seleccionadas) pueden sobresalir es en la coerción. Un Estado puede utilizar métodos de selección de la motivación para asegurarse de que sus policías, militares, servicios de inteligencia, y administración civil son leales de manera uniforme. Como señala Shulman,

> Los estados guardados [de alguna emulación fiel que se hubiera preparado y verificado cuidadosamente] se podrían copiar miles de millones de veces para conformar ideológicamente de manera uniforme a una fuerza militar burocrática y policial. Después de un corto período de trabajo, cada copia sería reemplazada por una copia nueva del mismo estado guardado, evitando una deriva ideológica. Dentro de una jurisdicción determinada, esta capacidad podría permitir una observación y regulación increíblemente detallada: podría haber una copia por cada residente. Esto podría ser utilizado para prohibir el desarrollo de armas de destrucción masiva, para hacer cumplir las regulaciones sobre la experimentación o reproducción de emulaciones cerebrales, o para imponer una constitución democrática liberal, o para crear un atroz y permanente totalitarismo.[39]

El primer efecto de una capacidad tal probablemente sería consolidar su poder y, posiblemente, concentrarlo en el menor número de manos.

Unificación por tratado

Puede haber grandes ganancias potenciales en colaborar a nivel internacional en un mundo multipolar post-transición. Las guerras y las carreras armamentísticas podrían evitarse. Los recursos astrofísicos podrían ser colonizados y ser cosechados a un ritmo global óptimo. El desarrollo de formas más avanzadas de inteligencia artificial podría coordinarse para evitar una carrera y permitir que los nuevos diseños fueran revisados a fondo. Otras novedades que pudieran representar riesgos existenciales podrían ser pospuestas. Y se podrían aplicar regulaciones uniformes a nivel mundial, incluyendo provisiones para un nivel garantizado de vida (lo que requeriría alguna forma de control de la población), y la prevención de la explotación y el abuso de emulaciones y otras mentes digitales y biológicas. Además, los

agentes con preferencias por recursos agotables (más sobre esto en el capítulo 13) preferirían un acuerdo de reparto que les garantizara una cierta porción frente a un futuro de lucha en la que un ganador se lo lleva todo y en la que correrían el riesgo de no conseguir nada.

La presencia de grandes ganancias potenciales en la colaboración, sin embargo, no implica que la colaboración realmente vaya a alcanzarse. En el mundo de hoy, muchas cosas grandes podrían obtenerse a través de una mejor coordinación global —la reducción de los gastos militares, de las guerras, de la sobrepesca, de las barreras comerciales y de la contaminación atmosférica, entre otras. Sin embargo, estos frutos carnosos se echan a perder en la rama. ¿Por qué sucede esto? ¿Qué impide un resultado plenamente cooperativo en el que se maximizara el bien común?

Uno de los obstáculos es la dificultad de garantizar el cumplimiento de cualquier tratado que pudiera acordarse, incluyendo los costes de vigilancia y aplicación. Dos rivales nucleares podrían estar mejor si ambos renunciaran a sus bombas atómicas; sin embargo, incluso si pudieran llegar a un principio de acuerdo para llevar a cabo el desarme, no obstante, éste podría ser difícil de alcanzar debido a su mutuo miedo a que la otra parte hiciera trampas. Disipar este temor requeriría la creación de un mecanismo de verificación. Puede que se necesitaran inspectores que supervisaran la destrucción de los arsenales existentes, y que luego controlaran los reactores nucleares y otras instalaciones, y que reunieran información de inteligencia técnica y humana, con el fin de asegurar que el programa de armas no se reconstituyera. Uno de los costes sería pagar a estos inspectores. Otro coste es el riesgo de que los inspectores espíen y trafiquen con secretos comerciales o militares. Y quizá lo más importante, que cada parte podría temer que la otra preservara su capacidad nuclear de manera clandestina. En muchos casos, un acuerdo potencialmente beneficioso no funciona porque el cumplimiento sería demasiado difícil de verificar.

Si llegaran a estar disponibles nuevas tecnologías de inspección que redujeran los costes de monitorización, podría esperarse que tuviera lugar a una mayor cooperación. Sin embargo, que los costes de monitorización fueran a reducirse en términos netos en la era post-transición, no está del todo claro. Si bien podría haber muchas nuevas y poderosas técnicas de inspección, también habría nuevos medios de ocultación. En particular, una parte creciente de las actividades que uno podría desear regular tendrían lugar en el ciberespacio, fuera del alcance de la vigilancia física. Por ejemplo, las mentes digitales que trabajen en el diseño de un nuevo sistema de armas nanotecnológicas o de una nueva generación de inteligencia artificial pueden hacerlo sin dejar ninguna huella física. El análisis forense digital puede no

penetrar todas las capas de ocultación y encriptado con las que un criminal encubriría sus actividades ilícitas.

Los detectores de mentiras fiables, si pudieran ser desarrollados, serían una herramienta muy útil para seguir el cumplimiento.[40] Un protocolo de inspección podría incluir disposiciones para entrevistar a funcionarios clave, con el fin de verificar que tienen intención de aplicar todas las disposiciones del tratado y que saben que no hay violaciones a pesar de haber hecho grandes esfuerzos para encontrarlas. Un tomador de decisiones que planeara engañar podría derrotar a un sistema de verificación basado en un detector de mentiras dando primero órdenes a sus subordinados para que llevaran a cabo la actividad ilícita y para que ocultaran la actividad al propio tomador de decisiones, y luego someterse a sí mismo a algún procedimiento que borrara el recuerdo de haber participado en estas maquinaciones. Certeras operaciones de borrado de memoria bien podrían ser factibles en cerebros biológicos con neurotec- nología más avanzada. Puede que esto sea aún más fácil en inteligencias artificiales (dependiendo de su arquitectura).

Los Estados podrían tratar de superar este problema comprometiéndose a un programa de monitorización permanente que pusiera a prueba regularmente a funcionarios clave con un detector de mentiras para comprobar si albergan alguna intención de subvertir o eludir cualquier tratado en el que haya entrado o pueda entrar el Estado en el futuro. Ese compromiso podría ser visto como una especie de meta-tratado, lo que facilitaría la verificación de otros tratados; pero los Estados pueden comprometerse a ello de forma unilateral para obtener el beneficio de ser considerado como un socio confiable en las negociaciones. Sin embargo, este compromiso o meta-tratado se enfrentaría al mismo problema de la subversión a través de una estratagema de delegación y olvido. Idealmente, el meta-tratado se pondría en vigor *antes* de que cualquiera de las partes tuviera la oportunidad de hacer los arreglos internos necesarios para subvertir su implementación. Una vez la maldad haya tenido un momento de descuido para sembrar sus minas de engaño, la confianza nunca puede pisar allí de nuevo.

En algunos casos, la mera capacidad de *detectar* violaciones de tratados es suficiente para establecer la confianza necesaria para un acuerdo. En otros casos, sin embargo, se necesita algún mecanismo para hacer cumplir o impartir castigo si se produce una violación. La necesidad de un mecanismo de *aplicación* puede surgir si la amenaza por parte de la parte agraviada de retirarse del tratado no es suficiente para disuadir las violaciones, por ejemplo, si el violador de tratados ganara una ventaja tal que dejara de preocuparse de la respuesta de la otra parte.

Si estuvieran disponibles métodos de selección de motivación altamente efectivos, este problema de aplicación podría ser resuelto mediante la potenciación de una agencia independiente con una fuerza policial o militar suficiente para hacer cumplir el tratado incluso en contra de la oposición de uno o varios de sus firmantes. Esta solución requiere que se pueda confiar en la agencia de aplicación. Pero con técnicas de selección de motivación lo suficientemente buenas, la confianza requerida podría lograrse haciendo que todas las partes del tratado supervisaran conjuntamente la designación de la agencia de aplicación.

Entregar el poder a un organismo de aplicación externa plantea muchos de los mismos problemas a los que nos enfrentamos antes en nuestros debates sobre un resultado unipolar (uno en el que una Unidad surgiera antes o durante la revolución inicial de la inteligencia artificial). Con el fin de ser capaz de hacer cumplir los tratados relativos a los vitales intereses de seguridad de los Estados rivales, la agencia de aplicación externa necesitaría, en efecto, constituirse como una Unidad: un Leviatán superinteligente global. Una diferencia, sin embargo, es que ahora estamos considerando una situación posterior a la transición, en la que los agentes que crearan este Leviatán serían más competentes que los seres humanos de la actualidad. Estos creadores del Leviatán podrían ser ellos mismos ya superinteligentes. Esto aumentaría en gran medida las probabilidades de que pudieran resolver el problema de control y diseño de una agencia de aplicación que sirviera a los intereses de todas las partes implicadas en su construcción.

Aparte de los costes de monitorización y de imposición del cumplimiento, ¿existen otros obstáculos para la coordinación global? Tal vez la cuestión pendiente importante es eso a lo que podríamos referimos como los *costes de negociación.*[41] Incluso cuando hay una posible negociación que beneficie a todos los involucrados, ésta a veces no llega a despegar porque las partes no logran ponerse de acuerdo sobre cómo dividir el botín. Por ejemplo, si dos personas pudieran hacer un acuerdo que produjera un resultado neto de un dólar, pero cada parte sintiera que merece sesenta centavos y se negara a conformarse con menos, el acuerdo no sucedería y la ganancia potencial se perdería. En general, las negociaciones pueden ser difíciles o prolongadas, o quedar totalmente estériles, debido a las estrategias de negociación tomadas por algunas de las partes.

En la vida real, los seres humanos con frecuencia tienen éxito llegando a acuerdos pese la posibilidad de negociación estratégica (aunque a menudo no sin un considerable gasto de tiempo y paciencia). Es concebible, sin embargo, que los problemas de negociación estratégica tuvieran una dinámica diferente en la era

post-transi- ción. Una IA negociadora podría adherirse de manera más consistente a cierta concepción particular y más formal de racionalidad, posiblemente con consecuencias nuevas o inesperadas cuando se combine con otras IAs negociadoras. Una IA también podría tener a su disposición movimientos del juego de negociación que, o no están disponibles para los seres humanos, o son mucho más difíciles de ejecutar para ellos, incluyendo la capacidad de comprometerse de antemano a una política o a un curso de acción. Mientras que los seres humanos (y las instituciones dirigidas por humanos) son en ocasiones capaces de comprometerse de antemano con grados imperfectos de credibilidad y especificidad —algunos tipos de inteligencia artificial podrían ser capaces de realizar arbitrarios compromisos previos irrompibles y permitir que los socios de negociación confirmaran que tal compromiso previo había sido hecho.[42]

La disponibilidad de técnicas de compromiso previo poderosas podría alterar profundamente la naturaleza de las negociaciones, lo que podría dar una ventaja inmensa a un agente que tuviera la ventaja del primer movimiento. Si es necesaria la participación de un agente particular para la realización de algunas ganancias potenciales en cooperación, y si ese agente es capaz de dar el primer paso, estaría en condiciones de dictar el reparto del botín comprometiéndose previamente a no aceptar ningún acuerdo que le diera menos de, digamos, el 99% de la plusvalía. Otros agentes se enfrentarían entonces con la opción de no conseguir nada (al rechazar la propuesta injusta) o conseguir el 1% del valor (cediendo). Si el compromiso previo del agente primero en mover fuera públicamente verificable, sus interlocutores en las negociaciones podrían estar seguros de que aquéllas son las únicas dos opciones.

Para evitar ser explotados de esta manera, los agentes podrían comprometerse de antemano a rechazar el chantaje y declinar todas las ofertas injustas. Una vez que un compromiso previo como éste se hubiera hecho (y se hubiera publicitado con éxito), otros agentes no tendrían interés en hacer amenazas o en comprometerse de antemano a sólo aceptar ofertas inclinadas a su favor, porque sabrían que las amenazas podrían fallar y que las propuestas injustas serían rechazadas. Pero esto sólo demuestra una vez más que la ventaja la tendría quien hiciera el primer movimiento. El agente que moviera primero podría elegir si desea aprovechar su posición de fuerza sólo para disuadir a otros de tomar ventaja injusta, o para conseguir la parte del león de futuras ganancias.

Podría parecer que el situado en mejor lugar sería el agente que comenzara con un temperamento o un sistema de valores que lo hiciera impermeable a la extorsión o a cualquier oferta —de un acuerdo en el que su participación sea indispensable— en la que no

esté recibiendo casi todas las ganancias. Algunos humanos ya parecen poseer rasgos de personalidad correspondientes a los de un espíritu inflexible.[43] Una disposición muy inflexible, sin embargo, podría ser contraproducente en caso de que hubiera otros agentes que también se sintieran con derecho a más de lo que les corresponde y estuvieran comprometidos a no dar marcha atrás. La fuerza imparable se encontraría entonces con el objeto inamovible, lo que resultaría en una falta de acuerdo (o peor: en la guerra total). Los mansos y los veleidosos conseguirían al menos algo, aunque menos de lo que les correspondería.

Qué tipo de equilibrio de teoría de juegos se alcanzaría en tal juego de negociación de post-transición no es inmediatamente evidente. Los agentes podrían elegir estrategias más complicadas que las que aquí se consideran. Uno *esperaría* que el equilibrio se alcanzara centrado en alguna norma de equidad que sirviera como punto de Sche- lling —una característica sobresaliente en una gama amplia de resultados que, debido a las expectativas compartidas, se convierte en un punto de coordinación probable en un juego de coordinación que de otro modo estaría indeterminado. Tal equilibrio puede ser reforzado por algunas de nuestras disposiciones evolucionadas y por programación cultural: una preferencia común para la equidad podría, suponiendo que tengamos éxito en la transferencia de nuestros valores en la era de post-transición, sesgar las expectativas y las estrategias de manera que condujeran a un equilibrio atractivo.[44]

En cualquier caso, el resultado es que, con la posibilidad de formas fuertes y flexibles de compromiso previo, los resultados de las negociaciones podrían derivar en un panorama desconocido. Aunque la era post-transición comenzara de manera multi- polar, podría convertirse en una Unidad casi de inmediato como consecuencia de un tratado que resolviera todos los problemas importantes de coordinación global. Algunos costes de transacción, incluyendo tal vez los costes de vigilancia y aplicación, podrían caer en picado con las nuevas capacidades tecnológicas disponibles para las inteligencias artificiales avanzadas. Otros gastos, en particular los costes relacionados con la negociación estratégica, podrían seguir siendo significativos. Pero de cualquier modo en que la negociación estratégica afecte a la naturaleza del acuerdo al que se llegue, no habría ninguna razón clara por la que debería prolongarse mucho el logro de un acuerdo —si dicho acuerdo fuera a alcanzarse en algún momento. Si no se alcanzara un acuerdo, entonces podría tener lugar algún tipo de lucha; en ese caso, una facción podría ganar y formar una Unidad en torno a la coalición ganadora, o el resultado podría ser un conflicto interminable, en cuyo caso la Unidad nunca se formaría y el resultado global quedaría muy lejos de lo que podría y debería

haberse alcanzado si la humanidad y sus descendientes hubieran actuado de manera más coordinada y cooperativa.

* * *

Hemos visto que la multipolaridad, incluso si se pudiera lograr de forma estable, no garantizaría un resultado atractivo. El problema del agente principal original permanecería sin resolver, y enterrarlo bajo un nuevo conjunto de problemas relacionados con fallos de coordinación globales de post-transición puede que sólo empeorara las cosas. Volvamos, pues, a la cuestión de cómo podemos mantener a raya a una única IA superinteligente.

CAPÍTULO 12

Adquiriendo valores

El control de la capacidad es, a lo sumo, una medida temporal y auxiliar. A menos que el plan sea mantener a la superinteligencia embotellada para siempre, será necesario dominar la selección de motivación. Pero, ¿cómo podríamos introducir valores en un agente artificial, con el fin de que éste persiguiera ese valor como su meta final? Mientras el agente no sea inteligente, puede ser que carezca de la capacidad de entender o incluso de representarse cualquier valor humanamente significativo. Sin embargo, si retrasamos el procedimiento hasta que el agente sea superinteligente, podría ser capaz de resistir nuestro intento de interferir con su sistema de motivación —y, como vimos en el capítulo 7, tendría razones instrumentales convergentes para hacerlo. Este problema de introducción de valores es difícil, pero debe ser afrontado.

El problema de la introducción de valores

Es imposible enumerar todas las situaciones posibles en que una superinteligencia se podría encontrar y especificar para cada situación la acción que debería tomar. Del mismo modo, es imposible crear una lista de todos los mundos posibles y asignar a cada uno de ellos un valor. En cualquier ámbito mucho más complicado que el juego de las tres en raya, hay demasiados estados posibles (e historias de estados) como para que la enumeración exhaustiva sea factible. Un sistema de motivación, por lo tanto, no se puede especificar como una tabla de búsqueda completa. En su lugar, debe expresarse de manera más abstracta, como una fórmula o regla que permita al agente decidir qué hacer en cualquier situación determinada.

Una manera formal de especificar una regla de decisión de este tipo es a través de una función de utilidad. Una función de utilidad (como se recuerda en el capítulo 1) asigna valor a cada resultado que podría obtenerse, o más en general, a cada "mundo posible". Dada una función de utilidad, se puede definir un agente que maximice la utilidad esperada. Tal agente seleccionaría en cada momento la acción que tuviera la utilidad esperada más alta. (La utilidad esperada se calcula ponderando la utilidad de cada mundo posible con la probabilidad subjetiva de que la realidad de ese mundo esté condicionada a que se elija una acción en particular). En realidad, los resultados posibles son demasiado numerosos como para poder calcular con exactitud la utilidad esperada de una acción. Sin embargo, la regla de la decisión y la función de utilidad en conjunto determinan una normativa ideal —una noción de optimización— en torno a la cual un agente podría ser diseñado para aproximarse; y la aproximación podría ir haciéndose más cercana a medida que el agente se volviera más inteligente.[1] La creación de una máquina que pudiera calcular una buena aproximación de utilidad esperada por las acciones disponibles para esta máquina es un problema de IA completa.[2] Este capítulo aborda otro problema, un problema que persistiera incluso si el problema de hacer máquinas inteligentes estuviera resuelto.

Podemos utilizar este marco teórico de un agente maximizador de utilidad para considerar la situación de un futuro programador de IA seminal que tuviera la intención de resolver el problema de control dotando a la IA con un objetivo final que correspondiera a cierta noción humanamente plausible de lo que fuera un resultado valioso. El programador tiene en mente algún valor humano en particular que le gustaría que la IA promoviera. Para ser concretos, vamos a decir la felicidad. (Cuestiones similares se plantean si el programador estuviera interesado en la justicia, la libertad, la gloria, los derechos humanos, la democracia, el equilibrio ecológico o el desarrollo personal). En términos del marco de utilidad esperada, el programador

busca así una función de utilidad que asigne utilidad a los mundos posibles en proporción a la cantidad de felicidad que contienen. Pero ¿cómo podría él expresar tal función de utilidad en código computacional? Los lenguajes de programación no contienen términos primitivos como "felicidad" Si tal término se fuera a utilizar, primero debería ser definido. No es suficiente definirlo en términos de otros conceptos humanos de alto nivel —"la felicidad es disfrutar de las potencialidades inherentes a nuestra naturaleza humana" o alguna paráfrasis filosófica similar. La definición debe basarse en términos que aparezcan en el lenguaje de programación de la IA, y, en última instancia, en términos primitivos tales como operadores matemáticos y direcciones que apunten a contenidos de registros de memoria individual. Cuando uno considera el problema desde esta perspectiva, se puede empezar a apreciar la dificultad de la tarea del programador.

Identificar y codificar nuestros propios objetivos finales es difícil porque las representaciones finales humanas son complejas. Sin embargo, debido a que la complejidad es en gran parte transparente para nosotros, a menudo no nos damos cuenta de que está ahí. Podemos comparar este caso con la percepción visual. La visión, del mismo modo, puede parecer una cosa simple, porque lo hacemos sin esfuerzo.[3] Sólo tenemos que abrir nuestros ojos, así nos parece, y una vista rica, significativa, eidéti- ca y tridimensional del entorno inunda nuestras mentes. Esta comprensión intuitiva de la visión es como la idea que tiene un duque de su hogar patriarcal: por lo que a él se refiere, las cosas simplemente aparecen en los momentos y lugares adecuados, mientras que los mecanismos que producen esas manifestaciones están ocultos a la vista. Sin embargo, lograr incluso la más simple tarea visual —encontrar la pimienta en la cocina— requiere de una enorme cantidad de trabajo computacional. A partir de una serie temporal confusa de patrones bidimensionales producidos por nervios excitándose, originados en la retina y transmitidos al cerebro a través del nervio óptico, la corteza visual debe trabajar retrospectivamente para reconstruir una representación tridimensional interpretativa del espacio exterior. Una parte importante de nuestro preciado metro cuadrado de tejido cortical se divide en zonas para el procesamiento de la información visual, y mientras usted está leyendo este libro, miles de millones de neuronas están trabajando sin descanso para lograr esta tarea (como si muchas costureras, inclinadas sobre sus máquinas de coser en una fábrica de explotación, estuvieran cosiendo y volviendo a coser una colcha gigante muchas veces por segundo). De la misma manera, nuestros valores y deseos aparentemente simples en realidad contienen una inmensa complejidad.[4] ¿Cómo pudo nuestro programador trasladar esta complejidad a una función de utilidad?

Un enfoque consistiría en tratar de codificar directamente una representación completa de cualquier meta que quisiéramos que la IA persiguiera; en otras palabras, escribir una función de utilidad explícita. Este enfoque podría funcionar si tuviéramos objetivos extraordinariamente simples, como por ejemplo, si quisiéramos calcular los dígitos de Pi —es decir, si lo único que quisiéramos fuera que la IA calculara los dígitos de Pi y fuéramos indiferentes a cualquier otra consecuencia que diera como resultado la búsqueda de esta meta— recuérdese nuestra discusión anterior sobre el modo de fallo de profusión infraestructural. Este enfoque de codificación explícita también podría ser prometedor en el uso de métodos de selección de motivación domésticos. Pero si tratamos de promover o proteger cualquier plausible valor humano, y estuviéramos construyendo un sistema destinado a convertirse en un soberano superinteligente, entonces la codificación explícita de la requerida representación completa de la meta parece estar irremediablemente fuera de nuestro alcance.[5]

Si no podemos transferir los valores humanos a una IA escribiendo representaciones completas en código computacional, ¿qué otra cosa podríamos probar? Este capítulo aborda varios caminos alternativos. Algunos de estos pueden parecer plausibles a primera vista, pero mucho menos al examinarlos detenidamente. Las futuras investigaciones deberían centrarse en los caminos que permanecen abiertos.

La solución al problema de introducción de valores es un reto de investigación digno de algunos de los mejores talentos matemáticos de la siguiente generación. No podemos posponer el enfrentamiento con este problema hasta que la IA haya desarrollado la suficiente capacidad de raciocinio como para entender fácilmente nuestras intenciones. Como vimos en el apartado sobre razones instrumentales convergentes, un sistema genérico resistirá los intentos de alterar sus valores finales. Si un agente no es ya fundamentalmente amigable en el momento en que adquiera la capacidad de reflexionar sobre su propia agencia, no verá con buenos ojos un intento tardío de lavado de cerebro o un complot para reemplazarlo por un agente diferente que ama mejor a su vecino.

Selección evolutiva

La evolución ha producido un organismo con valores humanos al menos una vez. Este hecho podría alentar la creencia de que los métodos evolutivos son el camino para resolver el problema de introducción de valores. Hay, sin embargo, graves obstáculos para

lograr seguridad a lo largo de este camino. Ya hemos señalado estos obstáculos al final del capítulo 10 cuando discutimos cómo los procesos de búsqueda poderosos pueden ser peligrosos.

La evolución puede ser vista como una clase particular de algoritmos de búsqueda que implican la alternancia de dos pasos: uno de expansión de la población de candidatos posibles mediante la generación de nuevos candidatos de acuerdo con alguna regla estocástica relativamente simple (por ejemplo, la mutación al azar o la recombinación sexual); el otro, contrayendo la población mediante la poda de candidatos que puntúan mal cuando se les prueba mediante una función de evaluación. Al igual que con muchos otros tipos de búsqueda de gran alcance, existe el riesgo de que el proceso encuentre una solución que satisfaga los criterios de búsqueda especificados formalmente pero no nuestras expectativas implícitas. (Esto se mantendría así tanto si buscáramos desarrollar una mente digital que tuviera los mismos objetivos y valores que un ser humano normal, como si en lugar de una mente buscáramos, por ejemplo, la perfección moral o la obediencia perfecta). El riesgo se evitaría si pudiéramos especificar un criterio de búsqueda formal que representa con precisión todas las dimensiones de nuestras metas, en lugar de sólo un aspecto de lo que pensamos que deseamos. Pero ése es precisamente el problema de la introducción de valores, y sería, por supuesto, una petición de principio en este contexto dar ese problema por resuelto.

Hay un problema adicional:

> La cantidad total de sufrimiento por año en el mundo natural está más allá de toda contemplación decente. Durante el minuto que me lleva a componer esta frase, miles de animales están siendo comidos vivos, otros están corriendo para salvar la vida, gimiendo de miedo, otros están siendo lentamente devorados desde dentro por parásitos, miles de todo tipo están muriendo de hambre, sed y enfermedad.[6]

Incluso sólo dentro de nuestra especie, 150.000 personas mueren cada día, mientras que muchas más sufren una serie de tormentos y carencias atroces.[7] La naturaleza podrá ser una gran experimentadora, pero nunca aprobaría un examen de ética —pues contraviene la declaración de Helsinki y todas las normas de decencia moral en todos los sentidos. Es importante que no repliquemos gratuitamente tales horrores *in silico*. Los crímenes mentales parecen especialmente difícil de evitar cuando se utilizan métodos evolutivos para producir inteligencia similar a la humana, al menos si el proceso está destinado a parecerse en algo a evolución biológica real.[8]

Aprendizaje por refuerzo

El aprendizaje por refuerzo es un área del aprendizaje artificial que

estudia técnicas mediante las cuales los agentes pueden aprender a maximizar una noción de recompensa acumulada. Mediante la construcción de un entorno en el que el rendimiento deseado sea recompensado, un agente de aprendizaje por refuerzo puede llegar a aprender a resolver una amplia clase de problemas (incluso en la ausencia de la instrucción detallada o la retroalimentación de los programadores, más allá de la señal de recompensa). A menudo, el algoritmo de aprendizaje implica la construcción gradual de una especie de función de evaluación, que asigna valores a los estados, a pares de estado-acción o a normas. (Por ejemplo, un programa puede aprender a jugar al backgammon mediante aprendizaje por refuerzo que mejore gradualmente su evaluación de las posibles posiciones en el tablero). La función de evaluación, que se actualiza de forma continua a la luz de la experiencia, podría ser considerada como la incorporación de una forma de aprender sobre valores. Sin embargo, lo que se aprende no son nuevos valores finales, sino que cada vez se hacen más precisas *las estimaciones de los valores instrumentales* de alcanzar ciertos estados particulares (o de las decisiones concretas en estados particulares, o de seguir algunas normas particulares). En la medida en que pueda decirse de un agente de aprendizaje por refuerzo que

tiene un objetivo final, ese objetivo se mantiene constante: maximizar la recompensa futura. Y la recompensa consiste en percepciones especialmente designadas recibidos de su entorno.

Por lo tanto, el síndrome de pinchazo cerebral sigue siendo un resultado probable en cualquier agente de refuerzo que desarrollara un modelo mundial lo suficientemente sofisticado como para sugerir esta forma alternativa de maximizar la recom- pensa.[9]

Estas observaciones no implican que los métodos de aprendizaje por refuerzo nunca pudieran ser utilizados en una IA seminal segura, sólo que tendrían que estar subordinados a un sistema de motivación que estuviera organizado en torno al principio de maximización de la recompensa. Eso, sin embargo, requeriría que una solución al problema de la introducción de valores hubiera sido encontrado por otros medios que el aprendizaje por refuerzo.

Acumulación de valores por asociación

Podríamos entonces preguntarnos: si el problema de la introducción de valores es tan complicado, ¿cómo conseguimos nosotros adquirir valores?

Un posible modelo (muy simplificado) podría ser algo como lo siguiente. Comenzamos la vida con algunas preferencias de partida relativamente simples (por ejemplo, una aversión a estímulos nocivos), junto con un conjunto de disposiciones para adquirir preferencias adicionales en respuesta a diversas experiencias posibles (por ejemplo, podríamos estar dispuestos a formar una preferencia por los objetos y comportamientos que encontráramos valorados y recompensados en nuestra cultura). Tanto las preferencias de partida simples y las disposiciones son innatas, moldeadas por la selección natural y sexual a lo largo de grandes períodos de tiempo evolutivo. Sin embargo, las preferencias con las que acabamos de adultos dependen de los acontecimientos de nuestra vida. Gran parte de la información contenida en nuestros valores finales se adquiere de nuestras experiencias, no está precargada en nuestros genomas.

Por ejemplo, muchos de nosotros queremos a otra persona y por lo tanto ponemos un gran valor final en su bienestar. ¿Qué se requiere para representar un valor como éste? Muchos elementos están involucrados, pero consideremos sólo dos: una representación de "persona" y una representación de "bienestar" Estos conceptos no son codificados directamente en nuestro ADN. Más bien, el ADN contiene instrucciones para la construcción de un cerebro, que, cuando se coloca en un entorno humano típico, desarrollará en el transcurso de varios años un modelo sobre el mundo que incluirá los conceptos de

personas y de bienestar. Una vez formados, estos conceptos pueden ser usados para representar ciertos valores significativos. Pero tiene que haber algún mecanismo innato que conduzca a que los valores se formen en torno a estos conceptos, en lugar de en torno a otros conceptos adquiridos (como los de una maceta o un sacacorchos).

Los detalles de cómo funciona este mecanismo no se conocen bien. En los seres humanos, el mecanismo es probablemente complejo y heterogéneo. Es más fácil de entender el fenómeno si lo consideramos en una forma más rudimentaria, como la impronta filial en aves nidífugas, donde los pollitos recién nacidos adquieren un

seo de proximidad física respecto al objeto que presente un estímulo de movimiento adecuado durante el primer día posterior a su eclosión. El objeto particular del que el pollito quiera estar cerca depende de su experiencia; solamente la disposición general para la impronta está determinada genéticamente. Análogamente, Harry podría colocar un valor final en el bienestar de Sally; pero si la pareja no se hubiera llegado a conocer nunca, podría haberse enamorado de otra persona en su lugar, y sus valores finales habrían sido diferentes. La capacidad de nuestros genes para codificar la construcción de un mecanismo de adquisición de objetivos explica cómo llegamos a tener objetivos finales de gran complejidad informativa, mayores de los que podrían estar contenidos en el propio genoma.

En consecuencia, podemos considerar si podríamos construir el sistema de motivación de una inteligencia artificial bajo los mismos principios. Es decir, en lugar de especificar directamente valores complejos, ¿podríamos especificar algún mecanismo que condujera a la adquisición de esos valores cuando la IA interactuara con un ambiente adecuado?

Imitar el proceso de acumulación de valores que se produce en los seres humanos parece difícil. El importante mecanismo genético del ser humano es el producto de millones de años de trabajo de la evolución, un trabajo que podría ser difícil de recapitular. Además, el mecanismo presumiblemente se ajusta estrechamente a la arquitectura neurocognitiva humana, y, por lo tanto, no es aplicable a inteligencias artificiales que no sean emulaciones de cerebro completo. Y si estuvieran disponibles emulaciones de cerebro completo de suficiente fidelidad, parecería más fácil comenzar con un cerebro adulto que tuviera ya cargadas representaciones completas de algunos valores humanos.[10]

Tratar de poner en práctica un proceso de acumulación de valores que imite en gran medida al de la biología humana, por tanto, parece una línea poco prometedora en el ataque al problema de la introducción de valores. Pero ¿podríamos tal vez diseñar un mecanismo artificial sustitutivo más descarado que hiciera que una IA importara representaciones de alta fidelidad de valores complejos importantes en su sistema de objetivos? Para que esto tuviera éxito, puede que no fuera necesario dar a la IA exactamente las mismas disposiciones de evaluación que las de un ser humano biológico. Puede incluso que no fuera deseable como objetivo —la naturaleza humana, después de todo, es imperfecta y con demasiada frecuencia revela una tendencia al mal, que sería intolerable en cualquier sistema cercano a lograr una ventaja estratégica decisiva. Mejor, tal vez, sería aspirar a un sistema de motivación que se apartara de la norma humana de manera sistemática, por ejemplo teniendo una tendencia

más sólida hacia adquirir objetivos finales altruistas, compasivos, o elevados en formas que reconoceríamos como reflejo de un excepcional buen carácter si estuvieran presentes en una persona humana. Para contar como mejoras, sin embargo, tales desviaciones de la norma humana tendrían que apuntar en direcciones muy particulares en lugar de al azar; y que continuaran presuponiendo la existencia de un marco antropocéntri- co estable de referencia que proporcionara generalizaciones axiológicas humanamente significativas (a fin de evitar el tipo de suplantaciones perversas de descripciones de los objetivos superficialmente plausibles, como hemos examinado en el capítulo 8). Es una cuestión abierta si esto es factible.

Una cuestión aún más a propósito de la acumulación de valores asociativos es que la IA podría desactivar el mecanismo de acumulación. Como vimos en el capítulo 7, la integridad del sistema de objetivos es un valor instrumental convergente. Cuando la IA alcanzara una cierta etapa del desarrollo cognitivo podría comenzar a considerar la operación continuada del mecanismo de acumulación como una influencia corruptora.[11] Esto no es necesariamente algo negativo, pero debería tenerse cuidado en que el sellado del sistema de objetivos ocurriera en el momento oportuno, después de que hubieran surgido los valores adecuados, pero *antes* de que se hubieran sobrescrito por acumulaciones adicionales no deseadas.

Andamiaje motivacional

Otro enfoque sobre el problema de introducción de valores es lo que podemos llamar andamiaje motivacional. Se trata de dar a la IA seminal un sistema objetivo provisional, con objetivos finales relativamente simples que podamos representar por medio de codificación explícita o algún otro método factible. Una vez que la IA haya desarrollado facultades de representación más sofisticadas, reemplazamos este andamiaje de sistema de objetivos provisional con uno que tenga diferentes objetivos finales. Este sistema de objetivos posterior gobernaría entonces a la IA cuando se convirtiera en una superinteligencia en toda regla.

Debido a que, para la IA, los objetivos del andamiaje no serían sólo objetivos instrumentales sino también *finales,* podría esperarse que la IA resistiera su reemplazo (la integridad de contenido de los objetivos no deja de ser un valor instrumental convergente). Esto crea un peligro. Si la IA tuviera éxito en frustrar la sustitución de su andamiaje de objetivos, el método fallaría.

Para evitar este tipo de fallo, es necesario tomar precauciones. Por

ejemplo, los métodos de control de capacidad podrían aplicarse para limitar los poderes de la inteligencia artificial hasta que el sistema de motivación maduro se hubiera instalado. En particular, se podría tratar de mantener su desarrollo cognitivo en un nivel que fuera seguro, pero que le permitiera representarse los valores que queremos incluir en sus objetivos finales. Para ello, se podría tratar de atrofiar diferencialmente ciertos tipos de habilidades intelectuales, tales como las necesarias para la formulación de estrategias y maquinaciones maquiavélicas, permitiendo al mismo tiempo (aparentemente) que habilidades más inocuas se desarrollaran hasta un nivel más alto.

También podrían usarse métodos de selección de la motivación que propiciaran una relación más colaborativa entre la IA seminal y el equipo programador. Por ejemplo, se podría incluir en el andamiaje del sistema de motivación el objetivo de favorecer el sentido dado por los programadores, incluyendo el permiso para sustituir cualquiera de los objetivos actuales de la IA.[12] Otros andamiajes de objetivos podrían incluir el ser transparente a los programadores sobre sus valores y estrategias, y el desarrollo de una arquitectura que fuera fácil de entender para los programadores y que facilitara la posterior implementación de un objetivo final humanamente significativo, así como motivaciones de domesticidad (como limitar el uso de los recursos computacionales).

Se podría incluso imaginar dotar a la IA seminal de un único objetivo final consistente de sustituirse a sí misma con un objetivo final diferente, uno que hubiera sido especificado por los programadores sólo de forma implícita o indirectamente. Algunas de las cuestiones planteadas por el uso de un andamiaje de objetivos "auto-reemplazable" surgen también en el contexto del enfoque de aprendizaje de valores, que se discutirá en la siguiente subsección. Algunas cuestiones adicionales serán discutidas en el capítulo 13.

El enfoque del andamiaje motivacional no está exento de inconvenientes. Uno es que se correría el riesgo de que la IA pudiera llegar a ser demasiado poderosa mientras todavía se esté ejecutando en su sistema de objetivos provisional. Entonces podría frustrar los esfuerzos de los programadores humanos de instalar el sistema de objetivos finales (ya sea mediante una resistencia contundente o a través de una subversión tranquila). Los viejos objetivos finales podrían entonces permanecer a cargo a medida que la IA seminal se convirtiera en una superinteligencia en toda regla. Otra desventaja es que la instalación de los objetivos deseados en una IA de nivel humano no es necesariamente mucho más fácil que hacerlo en una IA más primitiva. Una IA de nivel humano es más compleja y podría haber desarrollado una arquitectura que fuera opaca y difícil de alterar. Una IA seminal, por el contrario, es como una *tabula rasa* en la

que los programadores podrían inscribir cualquier estructura que consideraran útil. Este inconveniente podría convertirse en algo positivo si se lograra dar a la IA seminal andamiajes de objetivos que le hicieran querer desarrollar una arquitectura que facilitara a los programadores sus esfuerzos posteriores para instalar los valores finales definitivos. Sin embargo, no está claro cómo de fácil sería dar una IA seminal andamiajes de objetivos con esta propiedad, y tampoco está claro cómo incluso una IA seminal muy bien motivada podría ser capaz de hacer un trabajo mucho mejor que un equipo de programación humano en el desarrollo de un buena arquitectura.

Aprendizaje de valores

Llegamos ahora a un importante pero sutil enfoque del problema de introducción de valores. Se trata de utilizar la inteligencia de la propia IA para que *aprenda* los valores que queremos que persiga. Para ello, debemos proporcionar un criterio para que la IA recoja, al menos implícitamente, un conjunto adecuado de valores. Podríamos entonces construir la IA para que actuara de acuerdo a sus mejores estimaciones de estos valores implícitamente definidos. Ella perfeccionaría continuamente sus estimaciones a medida que aprendiera más sobre el mundo y poco a poco desentrañara las implicaciones de los criterios determinadores de valor.

En contraste con el enfoque del andamiaje, que da a la IA un andamiaje de objetivos provisional y más tarde lo reemplaza con unos objetivos finales diferentes, el enfoque de aprendizaje de valores conserva un objetivo final que no cambia a lo largo de las fases de desarrollo y funcionamiento de la IA. El aprendizaje no cambia el objetivo. Cambia solamente las creencias de la IA acerca del objetivo.

La IA, por lo tanto, debe estar dotada de un criterio que pueda utilizar para determinar qué percepciones constituyen pruebas a favor de determinadas hipótesis sobre lo que el objetivo final es, y qué percepciones constituyen pruebas en su contra. La especificación de un criterio adecuado podría ser difícil. Parte de la dificultad, sin embargo, atañe primeramente al problema de la creación de una inteligencia artificial general fuerte, lo cual requiere un poderoso mecanismo de aprendizaje que pueda descubrir la estructura del entorno a partir de entradas sensoriales limitadas. Ese problema podemos dejarlo de lado por el momento. Pero incluso habiendo solucionado el problema de cómo crear una IA superinteligente, persisten las dificultades específicas del problema de introducción de valores. Con el enfoque de aprendizaje de valores, éstos pasan a consistir en la necesidad de definir un criterio que conecte estímulos

perceptivos con hipótesis sobre valores.

Antes de ahondar en los detalles de cómo se podría implementar el aprendizaje de valores, podría ser útil ilustrar la idea general con un ejemplo. Supongamos que escribimos la descripción de un conjunto de valores en una hoja de papel. Doblamos el papel y lo ponemos en un sobre cerrado. Creamos entonces un agente con inteligencia general de nivel humano y le damos el siguiente objetivo final: "Maximizar la realización de los valores descritos en el sobre". ¿Qué hará este agente?

El agente no sabe inicialmente lo que está escrito en el sobre. Pero puede formular hipótesis, y puede asignar a esas hipótesis una probabilidad en función de sus principios y de los datos empíricos disponibles. Por ejemplo, el agente podría haber encontrado otros ejemplos de textos de autoría humana, o podría haber observado algunos patrones generales de comportamiento humano. Esto le permitiría hacer conjeturas. No se necesita un título en psicología para predecir que lo más probable es que la nota describa un valor como "minimizar la injusticia y el sufrimiento innecesario" o "maximizar la rentabilidad para los accionistas" que un valor como "cubrir todos los lagos con bolsas de plástico".

Cuando el agente toma una decisión, busca realizar acciones que sean eficaces para la consecución de los valores que considera más probable que se describan en la carta. Es importante destacar que el agente daría un alto valor instrumental a aprender más acerca de lo que dice la carta. La razón es que para casi cualquier valor final que pueda ser descrito en la carta, es más probable que se realice ese valor si el agente comprende lo que es, ya que el agente luego perseguirá ese valor con mayor eficacia. El agente también descubriría las razones instrumentales convergentes descritas en el capítulo 7 — integridad del sistema de objetivos, mejora cognitiva, adquisición de recursos, y así sucesivamente. Sin embargo, en el supuesto de que el agente asignara una probabilidad suficientemente alta a que los valores descritos en la carta implicaran el bienestar humano, no perseguiría estos valores instrumentales de inmediato convirtiendo el planeta en computronium y, por tanto, exterminando la especie humana, ya que al hacerlo se arriesgaría a estar destruyendo para siempre su capacidad para comprender su valor final.

Podemos comparar este tipo de agente a una barcaza unida a varios remolcadores que tiran en direcciones diferentes. Cada remolcador correspondería a una hipótesis sobre valor final del agente. La potencia del motor de cada remolcador corresponde a la probabilidad de la hipótesis asociada, y por lo tanto cambiaría a medida que se presenten nuevas evidencias, produciendo ajustes en la dirección del movimiento de la barcaza. La fuerza resultante debe

mover la barcaza a lo largo de una trayectoria que facilite el aprendizaje del valor final (implícito), evitando las derivas hacia una destrucción irreversible; y más tarde, cuando se alcanzara el mar abierto del conocimiento preciso sobre el valor final, el remolcador que todavía ejerciera una fuerza significativa tiraría la barcaza hacia la realización del valor descubierto a lo largo de la ruta más directa o más propicia.

Las metáforas del sobre y la barcaza ilustran el principio subyacente al enfoque de aprendizaje de valores, pero obvian una serie de cuestiones técnicas críticas. Estás se verán claramente una vez que empecemos a desarrollar nuestro enfoque dentro de un marco formal (véase el cuadro 10).

lugar para cada posible función de utilidad *U,* calcular la probabilidad condicional de que *U* satisfaga el criterio de valor *V* (condicionado a que w sea el mundo real). En tercer lugar para cada posible función de utilidad U, calcular la utilidad del mundo posible w. En cuarto lugar, combinar estas cantidades para calcular la utilidad esperada de la acción w. En quinto lugar, repetir este procedimiento para cada acción posible y realizar la acción que se descubriera como la de mayor continuidad en utilidad esperada (utilizando algún método arbitrario para romper lazos). Como se ha descrito, este procedimiento —que implica una consideración explícita y separadamente de cada mundo posible— es, por supuesto, salvajemente intratable de manera computacional. La IA tendría que utilizar métodos abreviados de cálculo que aproximaran esta noción optimización.

La pregunta, entonces, sería cómo definir este criterio de valor V.[18] Una vez que la IA tuviera una representación adecuada del criterio de valor podría en principio utilizar su inteligencia general para recopilar información acerca de qué mundos posibles son de manera más probable el real. A continuación, podría aplicar el criterio, en cada uno de esos mundos posibles w, para saber qué función de utilidad satisfaría el criterio *V* en w. Así se podría considerar la fórmula IA-VL como una forma de identificar y separar este reto clave en el enfoque del aprendizaje de valores —el reto de cómo representar *V.* El formalismo también saca a la luz otra serie de cuestiones (por ejemplo, cómo definir Y, W y U), que tendrían que ser resueltas antes de que se pudiera hacer funcionar este enfoque.[19]

Una cuestión pendiente sería cómo dotar a la IA de un objetivo del tipo "Maximi- zar la realización de los valores descritos en el sobre". (En la terminología del cuadro 10, la forma de definir el criterio de valor V). Para ello, es necesario identificar el lugar en el que se describen los valores. En nuestro ejemplo, esto requiere hacer una referencia exitosa a la carta del sobre. Aunque esto puede parecer trivial, no deja de haber dificultades. Para mencionar sólo una: es fundamental que la referencia no sea simplemente un objeto físico externo, sino un objeto en un momento determinado. De lo contrario la IA puede determinar que la mejor manera de alcanzar su objetivo es sobrescribiendo la descripción original de valor con una que proporcione un objetivo más fácil (por ejemplo, el valor consistente en que para cada número entero haya un número entero más grande). Una vez hecho esto, la IA podría inclinarse en la silla y relajarse —aunque es más probable que sobreviniera un fallo maligno, por razones que hemos discutido en el capítulo 8. Así que ahora nos enfrentamos a la cuestión de cómo definir el tiempo. Podríamos señalar a un reloj y decir: "El tiempo se define por los movimientos de este dispositivo", pero esto podría fallar si la IA conjeturara que puede manipular el tiempo moviendo las manecillas del reloj, una conjetura que de hecho sería correcta si se le diera la anterior definición de "tiempo". (En un caso realista, las cosas se complican aún más por el hecho de que los valores pertinentes no se describirían

convenientemente en una carta; es más probable que tuvieran que ser inferidos a partir de observaciones de estructuras preexistentes que contuvieran implícitamente la información relevante, como los cerebros humanos).

Otra cuestión sobre la codificación del objetivo "Maximizar la realización de los valores descritos en el sobre" es que incluso si todos los valores correctos fueran des-

critos en una carta, e incluso si el sistema de motivación de la IA se hubiera introducido correctamente en esta fuente, la IA podría no interpretar las descripciones de la forma que queríamos. Esto crearía un riesgo de suplantación perversa, como se discute en el capítulo 8.

Para aclarar, la dificultad aquí no es tanto la forma de garantizar que la IA pueda entender las intenciones humanas. Una superinteligencia debería desarrollar fácilmente tal entendimiento. Más bien, la dificultad es asegurarse de que la IA se sintiera motivada a perseguir los valores descritos en la forma en que pretendíamos. Esto no estaría garantizado por la capacidad de la IA de entender nuestras intenciones: una IA podría saber exactamente lo que queríamos decir y, sin embargo, ser indiferente a esa interpretación de nuestras palabras (estando motivada por alguna otra interpretación de las palabras o siendo indiferente a nuestras palabras por completo).

La dificultad se agrava por el desideratum de que, por razones de seguridad, la motivación correcta debería idealmente ser instalada en la IA seminal antes de que fuera capaz de representarse plenamente los conceptos humanos o entender las intenciones humanas. Para ello sería necesario que de alguna manera se creara un marco cognitivo, con una localización particular dentro de ese marco designado en el sistema de motivación de la IA como depósito de su valor final. Pero el marco cognitivo en sí debe ser revisable, a fin de permitirle a la IA ampliar sus capacidades de representación a medida que aprenda más sobre el mundo y se haga más inteligente. La IA podría sufrir el equivalente a las revoluciones científicas, en las que su visión del mundo se ve sacudida y en la que tal vez sufra crisis ontológicas en la que descubra que sus maneras anteriores de pensar sobre los valores se basan en confusiones e ilusiones. Sin embargo, empezando en un nivel sub-humano del desarrollo y continuando a lo largo de todo su desarrollo posterior hasta una superinteligencia galáctica, el comportamiento de la IA sería guiado por un valor final esencialmente inmutable, un valor final que la IA iría entendiendo mejor como consecuencia directa de su progreso —y probablemente sería entendido de manera muy diferente por la IA madura a como era entendida por sus programadores originales, aunque no diferente de manera aleatoria u hostil, sino de una manera benigna y apropiada. Cómo lograr esto sigue siendo una cuestión abierta.[20] (véase el cuadro 11).

cada vez más por el contenido real de *F.* Por lo tanto, es de esperar que la IA se vuelva cada vez más agradable cuanto más aprenda y más inteligente se vuelva.

Los programadores pueden colaborar en este proceso, y reducir el riesgo de que la IA cometa algún error catastrófico mientras su comprensión de *F* sea aún incompleta, proporcionando a la IA con "afirmaciones de programador", hipótesis sobre la naturaleza y el contenido de *F* a las que en un principio se asigna una alta probabilidad. Por ejemplo, a la hipótesis de que "engañar a los programadores es poco amigable" se le puede dar una alta probabilidad a priori. Estas afirmaciones del programador sin embargo, no son "verdaderas por definición" —no son axiomas incuestionables sobre el concepto de amistad. Más bien, son hipótesis iniciales sobre la amistad, hipótesis a las que una IA racional asignará una alta probabilidad al menos durante el tiempo en que confíe más en las capacidades epistémicas de los programadores que en la propia capacidad.

La propuesta de Yudkowsky también implica el uso de lo que él llama "semánticas de validez causal". La idea aquí es que la IA no debería hacer exactamente lo que los programadores le dijeran hacer sino más bien (algo así como) lo que le estuvieran tratando de decir que hiciera. Mientras los programadores tratan de explicar a la IA seminal lo que es la amigabilidad, pueden cometer errores en sus explicaciones. Por otra parte, los propios programadores no pueden comprender plenamente la verdadera naturaleza de la amistad. Por lo tanto, sería deseable que la IA tuviera la capacidad de corregir errores del pensamiento de los programadores, y de inferir la verdad o la intención contenida en cualquiera de las explicaciones imperfectas que los programadores lograran proporcionar Por ejemplo, la IA debería ser capaz de representar los procesos causales mediante los cuales los programadores aprenden y se comunican acerca de la amistad. Por lo tanto, tomando un ejemplo trivial, la IA debería entender que hay posibilidades de que un programador pueda cometer un error mientras ingresa información acerca de la amistad, y que la IA debería entonces tratar de corregir el error En términos más generales, la IA debería tratar de corregir cualquier influencia distorsionadora que pudiera haber corrompido el flujo de información sobre la amistad mientras pasaba de su fuente a través de los programadores a la IA (donde "distorsionadora" es una categoría epistémica). De manera ideal, a medida que la IA madurara, debería ir superando los sesgos cognitivos y otros conceptos erróneos más fundamentales que impedían a sus programadores comprender plenamente qué es la amistad.

El enfoque “Hail Mary” requiere fe en que haya otras superinteligencias por ahí que compartan suficientemente nuestros valores.[25] Esto hace que el planteamiento no sea ideal. Sin embargo, los obstáculos técnicos a los que se enfrenta el enfoque de “Hail Mary”, aunque importantes, posiblemente serían menos formidables que a los que se enfrentan los enfoques alternativos. Explorando enfoques no ideales pero fácilmente realizables, puede tener sentido, no con la intención de utilizarlos, sino para tener algo de lo que echar mano en caso de que las soluciones ideales no estuvieran listas a tiempo.

Recientemente, Paul Christiano ha propuesto otra idea para resolver el problema de introducción de valores.[26] Como el “Hail Mary”, es un método de aprendizaje de valores que intenta definir el criterio de valor por medio de un “truco” en lugar de llevar a cabo una construcción laboriosa. Por contraste con el “Hail Mary”, no presupone la existencia de otros agentes superinteligentes que pudiéramos señalar como modelos a seguir para continuar nuestra propia IA. La propuesta de Christiano es difícil de explicar brevemente —incluye una serie de consideraciones arcanas— pero podemos tratar al menos de exponer sus principales elementos.

Supongamos que pudiéramos obtener (a) una especificación matemáticamente precisa de un cerebro humano particular y (b) un entorno virtual matemáticamente bien especificado que contuviera una computadora ideal con una enorme y arbitraria cantidad de memoria y potencia de CPU. Dados (a) y (b), podríamos definir una función de utilidad *U* como la productividad que un cerebro humano produciría después de interactuar con este entorno. *U* sería un objeto matemático bien definido, aunque uno que (debido a las limitaciones computacionales) podríamos ser incapaces de describir de forma *explícita.* Sin embargo, *U* podría servir como criterio de valor para un aprendizaje de valores de IA, que podría utilizar diversas heurísticas para asignar probabilidades a hipótesis sobre lo que *U* implica.

Intuitivamente, queremos que *U* sea la función de utilidad que un ser humano debidamente preparado produciría si tuviera la ventaja de poder utilizar una gran cantidad de potencia computacional —suficiente potencia computacional, por ejemplo, como para ejecutar números astronómicos de copias de sí mismo para que le ayudaran con su análisis en la especificación de una función de utilidad, o para ayudarle a diseñar una mejor manera de abordar este análisis. (Estamos aquí presagiando un tema, la “voluntad extrapolada coherente”, que será explorado en más profundidad en el capítulo 13).

Parecería relativamente fácil especificar el entorno ideal: podemos dar una descripción matemática de un ordenador abstracto con una capacidad arbitrariamente grande; y para otros aspectos podríamos utilizar un programa de realidad virtual que diera una descripción matemática de, por ejemplo, una habitación individual con un ordenador en ella (plasmando la computadora abstracta). Pero, ¿cómo obtener una descripción matemática precisa de un cerebro humano en particular? La forma más obvia sería a través de la emulación de cerebro completo, pero ¿qué ocurre si la tecnología de emulación no está disponible para entonces?

Aquí es donde la propuesta de Christiano ofrece una innovación clave. Christiano observa que con el fin de obtener un criterio de valor matemáticamente bien especi-

continúa

ficado, no necesitamos un modelo computacional prácticamente útil de una mente, un modelo que pudiéramos ejecutar. Sólo necesitaríamos una *definición* (posiblemente implícita y desesperadamente complicada) matemática —y esto podría ser mucho más fácil de alcanzar Con el uso de la neuroimagen funcional y otras mediciones, tal vez podamos recopilar gigabytes de datos sobre el comportamiento de entrada-salida de un humano seleccionado. Si recopiláramos una cantidad suficiente de datos, entonces podría ser que el modelo matemático más simple que diera cuenta de todos estos datos fuera, de hecho, una emulación del humano en cuestión. Aunque sería computacionalmente intratable para nosotros *encontrar* este modelo más simple a partir de los datos, podría ser perfectamente posible que nosotros *definiéramos* el modelo, haciendo referencia a los datos y usando una medida de simplicidad matemáticamente bien definida (como alguna variante de la complejidad de Kolmogorov, que nos encontramos en el cuadro 1, capítulo 1).[27]

En resumen, aún no se sabe cómo utilizar el enfoque de aprendizaje de valores para instalar valores humanos plausibles (aunque véase el cuadro 12 para ver algunos ejemplos de ideas recientes). En la actualidad, el enfoque debe ser visto como un programa de investigación en lugar de una técnica disponible. Si se pudiera hacerlo funcionar, podría constituir la mejor solución al problema de introducción de valores. Entre otros beneficios, parece ofrecer una forma natural de prevenir los crímenes mentales, ya que una IA seminal que hiciera conjeturas razonables sobre los valores que sus programadores podrían haberle instalado anticiparía que los crímenes mentales probablemente serían evaluados negativamente, y por lo tanto que sería mejor evitarlos, por lo menos hasta que hubiera obtenido información más concluyente.

Por último, pero no menos importante, está la cuestión de "qué escribir en el sobre" —o, menos metafóricamente, la cuestión de qué valores debemos tratar que la IA aprenda. Pero este problema es común a todos los enfoques del problema de introducción de valores en IA. Volveremos sobre ello en el capítulo 13.

Modulación de emulaciones

El problema de introducción de valores se presenta de manera diferente para la emulación de cerebro completo que para la inteligencia artificial. Los métodos que presuponen un fino conocimiento y control de los algoritmos y arquitecturas no se aplican a las emulaciones. Por otro lado, el método de selección de la motivación de la aumentación —inaplicable a la inteligencia artificial de novo— puede ser utilizado con emulaciones (o cerebros biológicos

mejorados).[28]

El método de aumentación podría combinarse con técnicas de ajuste de objetivos heredados del sistema. Por ejemplo, se podría tratar de manipular el estado de motivación de una emulación mediante la administración del equivalente digital de sustancias psicoactivas (o, en el caso de sistemas biológicos, de productos químicos reales). Incluso ahora es posible manipular farmacológicamente valores y motivaciones hasta un determinado punto.[29] La farmacopea del futuro puede contener fármacos

con efectos más específicos y predecibles. El medio digital de las emulaciones debería facilitar en gran medida esta evolución, haciendo que la experimentación controlada fuera más fácil, y haciendo posible manipular directamente todas las partes cerebrales.

Al igual que cuando se utilizan los sujetos de prueba biológicos, la investigación con emulaciones podría enredarse en complicaciones éticas, no todas las cuales podrían evitarse con un formulario de consentimiento. Tales enredos podrían frenar el progreso a lo largo del camino de la emulación (debido a la regulación o restricción moral), tal vez especialmente obstaculizando estudios sobre cómo manipular la estructura motivacional de las emulaciones. El resultado podría ser que las emulaciones crecieran hasta niveles superinteligentes de capacidad cognitiva potencialmente peligrosos antes de que el trabajo adecuado para probar o ajustar sus objetivos finales se hubiera hecho. Otro posible efecto de los enredos morales podría ser el de dar ventaja a los equipos y naciones menos escrupulosas. Por el contrario, si relajáramos nuestros estándares morales para experimentar con mentes humanas digitales, podríamos hacernos responsables de una cantidad sustancial de daño y maldad, lo cual es obviamente indeseable. Si los demás factores no se alteraran, estas consideraciones estarían a favor de tomar algún camino alternativo que no requiriera el uso extensivo de sujetos humanos de investigación digitales en una situación estratégica de alto riesgo.

La cuestión, sin embargo, no está clara. Se podría argumentar que la investigación sobre la emulación de cerebro completo es menos susceptible de incurrir en violaciones morales que la investigación sobre inteligencia artificial, sobre la base de que somos más propensos a reconocer el estatus moral de una mente de emulación que el estatus moral de una mente completamente extraña o sintética. Si ciertos tipos de IAs, o sus subprocesos, tuvieran un estatus moral significativo que no fuéramos capaces de reconocer, las violaciones morales consiguientes podrían ser extensas. Consideremos, por ejemplo, el feliz abandono con el que los programadores contemporáneos crean agentes de refuerzo del aprendizaje y los someten a estímulos aversivos. Innumerables agentes como éstos son creados todos los días, no sólo en los laboratorios de informática sino en muchas aplicaciones, incluyendo algunos juegos de ordenador que contienen sofisticados personajes no manejados por los jugadores. Presumiblemente, estos agentes son todavía demasiado primitivos como para tener alguna condición moral. ¿Pero cuánta confianza podemos tener de que esto es realmente así? Más importante aún, ¿cuánta confianza podemos tener en que vamos a saber parar a tiempo, antes de que nuestros programas sean capaces de experimentar un sufrimiento moralmente relevante?

(Volveremos en el capítulo 14 sobre algunas de las preguntas estratégicas más amplias que surgen cuando se compara la conveniencia de los caminos de la emulación y la inteligencia artificial).

Diseño institucional

Algunos sistemas inteligentes consisten en piezas inteligentes que son a su vez capaces de agencia. Las empresas y los Estados ejemplifican esto en el mundo humano: están en gran parte compuestos de seres humanos que pueden, para algunos propósitos, ser vistos como agentes autónomos en su propio derecho. Las motivaciones de estos sistemas compuestos no sólo dependen de las motivaciones de sus subagentes constituyentes, sino también de cómo se organizan los subagentes. Por ejemplo, un grupo que se organiza bajo una fuerte dictadura podría comportarse como si tuviera una voluntad idéntica a la voluntad del subagente que ocupa el papel del dictador, mientras que un grupo democrático a veces podría comportarse como si tuviera una voluntad que fuera un compuesto o promedio de las voluntades de sus diversos componentes. Pero también podríamos imaginar las instituciones de gobierno que harían que una organización se comportara de una manera que no fuera una simple función de las voluntades de sus subagentes. (En teoría, al menos, podría existir un Estado totalitario que todo el mundo odiara, si el Estado tuviera mecanismos que previnieran las revueltas entre sus ciudadanos. Cada ciudadano podría salir peor parado rebelándose que ocupando su lugar en la maquinaria del Estado).

Mediante el diseño de instituciones adecuadas para un sistema compuesto, podría tratarse de conformar su motivación efectiva. En el capítulo 9, hablamos de la integración social como un posible método de control de capacidad. Pero no nos hemos centrado en los incentivos a los que se enfrenta un agente como consecuencia de su existencia en un mundo social de casi-iguales. Aquí nos estamos centrando en lo que sucede *dentro* de un agente dado: cómo su voluntad está determinada por su organización interna. Estamos, por tanto, ante un método de selección de la motivación. Además, dado que este tipo de diseño institucional interno no depende de la ingeniería o reforma social a gran escala, es un método que podría estar disponible para un proyecto individual de desarrollo de la superinteligencia incluso si el entorno socio-económico o internacional en general no fuera idealmente favorable.

El diseño institucional es quizás más plausible en contextos en los que se combina con el de aumentación. Si pudiéramos comenzar con agentes que ya estuvieran debidamente motivados o que tuvieran

motivaciones similares a las humanas, los arreglos institucionales podrían ser utilizados como una salvaguardia adicional para aumentar las posibilidades de que el sistema mantuviera el rumbo.

Por ejemplo, supongamos que empezamos con algunos agentes bien motivados y parecidos a los humanos —por ejemplo, las emulaciones. Queremos potenciar las capacidades cognitivas de estos agentes, pero nos preocupa que las mejoras pudieran dañar sus motivaciones. Una manera de hacer frente a este reto sería la creación de un sistema en el que las emulaciones individuales funcionaran como subagentes. Cuando se introdujera una nueva mejora, inicialmente se aplicaría a un pequeño subconjunto de subagentes. Sus efectos serían luego estudiados por un panel de revisión formado por subagentes que aún no se les hubiera aplicado dicha mejora. Sólo cuando estos compañeros se hubieran cerciorado de que la mejora no fuera corruptora, la extenderían a una población de subagentes más amplia. Si encontraran que los subagentes mejorados se hubieran corrompido, no se les daría nuevas mejoras y se les excluiría de la toma de decisiones clave (por lo menos hasta que el sistema en su conjunto hubiera avanzado hasta un punto en que los subagentes corruptos pudieran reintegrarse de manera segura).[30] Aunque los subagentes corruptos podrían haber ganado alguna ventaja de la mejora, la estructura institucional en la que están inmersos, y el hecho de que constituyen una pequeña minoría entre los subagentes, probablemente les impediría tomar el poder o propagar su corrupción al sistema en general. Por lo tanto, la inteligencia colectiva y la capacidad del sistema se podría mejorar gradualmente en una secuencia de pequeños pasos, donde la solidez de cada paso sería verificada por subagentes sólo ligeramente menos capaces que los nuevos sub-agentes que hubieran dado ese paso.

La cantidad de la seguridad obtenible mediante este tipo de diseño institucional está limitada por la precisión de las pruebas que se utilicen para evaluar a los sub-agentes mejorados. El problema de la construcción de ensayos se complica por el hecho de que las baterías de pruebas sólo podrían ser validadas dentro del rango pre-existente de capacidades cognitivas pero tendrían que ser aplicadas a agentes que estarían fuera de este rango. Se podría mitigar este problema dando pequeños pasos de mejora y dejando que las pruebas duraran mucho tiempo. Tal precaución, sin embargo, podría elevar el coste y ralentizar el progreso (que, en caso de que estuviera teniendo lugar una carrera dinámica, podría significar que el proyecto que empleara estas medidas de seguridad se estaría poniendo en desventaja). Incluso con pequeños pasos y pruebas exhaustivas, quedaría algún riesgo de que se introdujeran defectos sutiles, lo que llevaría a una corrupción gradual del sistema objetivo, una corrupción que podría manifestarse

sólo cuando el sistema alcanzara la superinteligencia. Este riesgo podría reducirse dejando que la condición de prueba se extendiera a muchas medidas de mejora de la cognición, por lo que incluso los agentes que hubieran sido inicialmente certificados como sanos y que hubieran recibido varias mejoras posteriores seguirían siendo objetos de examen por un conjunto de compañeros originales sin modificar e incorruptos y quedarían sujetos a reversión.

Idealmente, la prueba debe ser continua en vez de limitarse a un período de evaluación inicial. Subagentes de nivel humano tendrían la capacidad de elaborar estrategias y, por lo tanto, podrían optar por ocultar ciertas metas mientras su comportamiento estuviera siendo monitorizado. Para que una prueba continua fuera posible, una jerarquía podría crearse en la que subagentes de un cierto nivel de capacidad tuvieran la tarea de monitorizar a subagentes de un nivel algo más elevado de capacidad. En la parte inferior de la jerarquía de capacidad (pero en la parte superior de la jerarquía de poder) se sentaría un director relativamente tonto y lento. Este director humano sería como un rey demente que reina sobre un tribunal incompetente que supervisa a una administración mediocre que gobierna a un pueblo capaz. Alternativamente, esto podría ser comparado con el evolutivamente antiguo "cerebro reptiliano", que se asegura de que el neocórtex, más nuevo y mucho más inteligente, sea empleado en servicio de objetivos tales como la alimentación y la cópula.

Uno podría preguntarse si una meritocracia inversa como ésta podría permanecer estable. Si resultara que la mayor parte de los subagentes superinteligentes fueran corruptos, y que los seres humanos a cargo de todo hubieran tratado de revertir muchas de las mejoras anteriores, ¿sería la estructura institucional lo suficientemente fuerte como para resistir la oposición que esta medida generaría? En el contexto de las sociedades humanas históricas o contemporáneas, probablemente la respuesta es que esto sería inverosímil. (No es que las personas más capaces se eleven constantemente al poder, pero tampoco estamos, *literalmente,* gobernados por chimpancés). La cuestión está menos clara, sin embargo, en un contexto donde la sociedad fuera implementada como un software.[31]

Imaginemos una jerarquía digital que tuviera un agente de policía altamente capacitado por cada dos proletarios. También contaría con un agente especial de policía por cada dos agentes de policía. También tendría un agente especial de la policía especial para cada dos agentes especiales de policía, —etc., de manera que cada capa de supervisores tuviera por lo menos la mitad de personas que la capa a la que supervisa. Imaginemos, además, que esta sociedad poseyera una arquitectura en la que el supervisor tuviera grandes ventajas sobre sus subordinados. Por ejemplo, el supervisor sería capaz de supervisar y

registrar todo lo que sus subordinados dicen y hacen, mientras que el supervisor mismo se ocultaría detrás de un vidrio de un solo sentido, de modo que sus subordinados sólo podrían oír las órdenes e instrucciones que los superiores decidieran transmitir. Imaginemos que los supervisores tuvieran un panel de botones que pudieran pulsar en cualquier momento para que un subordinado fuera castigado o recompensado, paralizado, eliminado o revertido a un estado anterior. Estas capacidades aumentarían considerablemente los poderes de un agente de supervisión, más allá de lo que ha sido históricamente posible incluso en las más totalitarias de las organizaciones.

Tampoco así se agotan los posibles instrumentos de control. Los supervisores también podrían tener a su disposición simulaciones de realidad virtual fidedignas que podrían utilizar para sondear las reacciones de los subordinados a escenarios hipotéticos (tales como simulaciones de ofertas para unirse a una conspiración o de vaguear cuando piensan que nadie les está mirando). Los supervisores pueden, además, tener la capacidad de espiar los monólogos internos de la mente de sus subordinados y manipular directamente sus estados emocionales.

El resultado sería una serie de ventajas enormes para el supervisor.[32] Cada supervisor sería un subordinado para otro supervisor de nivel superior, que vigilaría cada uno de *sus* movimientos y se aseguraría de que éste no se relajara en el ejercicio de sus funciones de gestión. Con este tipo de acuerdo, podría ser posible que la estructura fuera estable incluso si cada nivel de supervisión fuera intelectualmente algo más débil que el nivel que supervisados.[33] Y toda la estructura, lo que podría incluir muchos agentes altamente superinteligentes, podría ser controlada por un pequeño número de seres humanos con acceso a la raíz.[34]

El coste de una estructura de control interno sería significativo, pero no necesariamente prohibitivo. Por ejemplo, en el modelo en el que cada nivel de supervisión tuviera la mitad de miembros que la capa de abajo, la sobrecarga computacional adicional estaría limitada a un mero 100% de lo que cuesta la parte proletaria —menos si la capa de jefes tontos requiriera menos cálculos por subagente. La supervisión podría ser perfeccionada mediante el aumento de la proporción, o el coste rebajado por la disminución de la misma. El coste computacional extra podría ser un factor importante en un escenario de mercado competitivo (cf. capítulo 11), pero parece asequible en el contexto de un proyecto que no se enfrentara a una fuerte competencia inmediata. También habría un coste en términos de tiempo de desarrollo necesario para crear y probar estas funciones de supervisión. Un proyecto bien financiado podría reducir este coste mediante la paralelización del desarrollo de la estructura de control

con el desarrollo de la inteligencia artificial; pero la carga de trabajo adicional podría ser prohibitiva para los proyectos de menor envergadura y para los proyectos inmersos en una reñida carrera tecnológica.

Otro tipo de costes también merecen consideración: el riesgo de los crímenes metales que se cometerían en este tipo de estructuras.[35] Como se ha descrito, la institución suena como un horrible campo de trabajo de Corea del Norte. Pero hay maneras de al menos mitigar los problemas morales del funcionamiento de este tipo de instituciones, incluso si los subagentes contenidos en la institución fueran emulaciones con un estatus moral humano completo. Como mínimo, el sistema podría apoyarse en emulaciones voluntarias. Cada subagente podría tener la opción en cualquier momento de retirar su participación.[36] Las emulaciones terminadas podrían ser almacenadas en memorias, con el compromiso de ser reiniciadas bajo condiciones mucho más ideales una vez que la fase peligrosa de la explosión de inteligencia hubiera terminado. Mientras tanto, los subagentes que decidieran participar podrían ser alojados en entornos virtuales muy confortables y darles suficiente tiempo para dormir y recrearse. Estas medidas supondrían un coste, el cual debería ser manejable para un proyecto bien financiado en condiciones no competitivas. En una situación altamente competitiva, el coste puede ser inaccesible a menos que una empresa pueda estar seguro de que sus competidores incurrirían en el mismo coste.

En el ejemplo, hemos imaginado a los subagentes como emulaciones. Uno podría preguntarse, ¿el enfoque de diseño institucional requiere que los subagentes sean antropomórficos? ¿O es igualmente aplicable a sistemas compuestos por subagentes artificiales?

Nuestra primera impresión sobre este punto podría ser escéptica. Se observará que a pesar de nuestra abundante experiencia con agentes de apariencia humana, todavía no podemos predecir con precisión el arranque o los resultados de las revoluciones; la ciencia social puede, a lo sumo, describir algunas tendencias estadísticas.[37] Ya que no podemos predecir con fiabilidad la estabilidad de las estructuras sociales de los seres humanos ordinarios (de la que tenemos muchos datos), es tentador inferir que tenemos pocas esperanzas de ser muy precisos sobre las estructuras sociales estables de agentes cognitivamente mejorados similares a los humanos (sobre los cuales no tenemos datos), y que tenemos mucha menos esperanza de conseguirlo para agentes artificiales avanzados (que ni siquiera son similares a los agentes sobre los que tenemos datos).

Sin embargo, el asunto no es tan blanco o negro. Los seres humanos y los seres parecidos a los humanos son complejos; pero los

agentes artificiales podrían tener arquitecturas relativamente simples. Los agentes artificiales también podrían tener motivaciones simples y explícitamente caracterizadas. Además, los agentes digitales en general (ya sean emulaciones o inteligencias artificiales) son copiables: una posibilidad que puede revolucionar la administración, de igual modo que las piezas intercambiables revolucionaron la fabricación. Estas diferencias, junto con la oportunidad de trabajar con agentes que inicialmente fueran impotentes y crear estructuras institucionales que utilizaran las distintas medidas de control antes mencionadas, podrían combinarse para que fuera posible lograr resultados institucionales concretos —como un sistema que no se rebelara— más fiables que si se estuviera trabajando con seres humanos en condiciones históricas.

Pero, de nuevo, los agentes artificiales pueden carecer de muchos de los atributos que nos ayudan a predecir el comportamiento de los agentes de apariencia humana. Los agentes artificiales no necesitarían tener ninguna de las emociones sociales que se asocian al comportamiento humano, emociones como el miedo, el orgullo y el remordimiento. Los agentes artificiales tampoco necesitarían desarrollar apego hacia amigos y familiares. Tampoco les es necesario presentar la expresión corporal inconsciente que hace que sea difícil para los seres humanos ocultar nuestras intenciones. Estos déficits pueden desestabilizar las instituciones de agentes artificiales. Además, los agentes artificiales podrían ser capaces de hacer grandes saltos de rendimiento cognitivo como resultado de aparentemente pequeños cambios en sus algoritmos o arquitecturas. Los agentes artificiales despiadadamente optimizadores podrían estar dispuesto a tomar medidas extremas frente a las cuales los humanos se asustarían.[38] Y los agentes superinteligentes podrían mostrar una sorprendente capacidad para coordinarse con poca o ninguna comunicación (por ejemplo, mediante el modelado interno de todas las hipotéticas respuestas a diversas contingencias). Estas y otras diferencias podrían hacer que un súbito fracaso institucional fuera más probable, incluso en el engranaje de lo que parecerían ser métodos muy encorsetados de control social.

No está claro, por lo tanto, cómo de prometedor sería el enfoque de diseño institucional, y si tendría más posibilidades de funcionar con agentes antropomórficos o con agentes artificiales. Se podría pensar que la creación de una institución con pesos y contrapesos adecuados sólo podría aumentar la seguridad o, en todo caso, no reduciría la seguridad, de modo que a partir de una perspectiva de la mitigación del riesgo siempre sería mejor que se utilizara el método. Pero incluso esto no puede afirmarse con certeza. El enfoque añade piezas y complejidad, y, por lo tanto, también puede introducir nuevas formas

de que las cosas salgan mal que no existirían en el caso de un agente que no tuviera subagentes inteligentes como partes. Sin embargo, el diseño institucional es digno de mayor exploración.[39]

Sinopsis

La ingeniería de sistemas de objetivos no es aún una disciplina establecida. No se sabe actualmente cómo transferir los valores humanos a un ordenador digital, incluso contando con un nivel humano de inteligencia artificial. Tras investigar una serie de enfoques, hemos encontrado que algunos de ellos parecen callejones sin salida; si bien otros parecen prometedores y merecen ser explorados más. Un resumen se presenta en la tabla 12.

Tabla 12. ***Resumen de las técnicas de introducción de valores***

Representación	Puede ser prometedora como una forma de introducir valores de domesticidad. No parece prometedora como una manera de introducir valores más complejos.
Selección evolutiva	Menos prometedora. Una búsqueda de gran alcance podría encontrar un diseño que satisficiera los criterios de búsqueda formales, pero no nuestras intenciones. Además, si los diseños fueran evaluados ejecutándose —incluyendo diseños que ni siquiera cumplan con los criterios formales— se crearía un peligro adicional potencialmente grave. La evolución también hace que sea difícil evitar los crímenes mentales masivos, sobre todo si se tiene el objetivo de diseñar mentes similares a las humanas.
Aprendizaje por refuerzo	Una gama de diferentes métodos pueden ser utilizados para resolver "problemas de aprendizaje por refuerzo", pero por lo general implican la creación de un sistema que busque maximizar una señal de recompensa. Esto tiene una tendencia inherente a producir el modo de fallo de pinchazo cerebral cuando el sistema se vuelva más inteligente. El aprendizaje por refuerzo, por lo tanto, parece poco prometedor

Tabla 12. *Continúa*

Acumulación de valores	Nosotros los humanos adquirimos gran parte del contenido específico de nuestros objetivos de reacciones a la experiencia. Mientras que la acumulación de valores podría, en principio, ser utilizado para crear un agente con motivaciones humanas, las disposiciones de acumulación de valores humanos podrían ser complejas y difíciles de replicar en una IA seminal. Una mala aproximación podría producir una IA que generalizara de manera diferente a los humanos y que por lo tanto adquiriera objetivos finales no deseados. Se necesita más investigación para determinar lo difícil que sería hacer el trabajo de acumulación de valores con suficiente precisión.
Anticipación de motivación	Es claro que no tenemos para decir lo difícil que sería fomentar que un sistema desarrollara representaciones internas de alto nivel que fueran transparentes para los seres humanos (manteniendo las capacidades del sistema por debajo de un nivel peligroso) y luego utilizar esas representaciones para diseñar un nuevo sistema de objetivos. Este enfoque podría ser considerablemente prometedor (Sin embargo, como con cualquier método no probado que aplace gran parte de la fuerza de trabajo en ingeniería de seguridad hasta el desarrollo de IAs de nivel humano, se debe tener cuidado de no permitir que se convierta en excusa para una actitud displicente frente al problema de control provisional).
Aprendizaje de valores	Un enfoque de valores es ciertamente prometedor pero es necesaria más investigación para

determinar lo difícil que sería especificar formalmente una referencia que apuntara con éxito a la información externa relevante sobre valores humanos (y de lo difícil que sería especificar un criterio de corrección para una función de utilidad en términos de dicha referencia). También valdría la pena explorar dentro de la categoría de aprendizaje de valores las propuestas del tipo "Hail Mary" o las que van en la línea de la construcción de Paul Christiano (u otros accesos directos).

Modulación de la inteligencia

Si se lograra la inteligencia artificial a través de la vía de la emulación, probablemente sería posible modificar las motivaciones a través del equivalente digital a drogas o similares. Que esto permitiera introducir valores con suficiente precisión como para garantizar la seguridad incluso cuando la emulación se hubiera elevado hasta la superinteligencia, es una pregunta abierta. (Las limitaciones éticas podrían complicar también la evolución en esta dirección).

Diseño institucional

Varios institutos métodos de control social podrían aplicarse en una institución compuesta por emulaciones. En principio, los métodos de control social también podrían aplicarse a una institución compuesta por inteligencias artificiales. Las emulaciones tienen algunas propiedades que hacen que sean más fáciles de controlar a través de estos métodos, pero también algunas propiedades que podrían hacer que fueran más difícil de controlar que las IAs. El diseño institucional parece digno de mayor exploración como una potencial técnica de introducción de valores.

Si supiéramos cómo resolver el problema de la introducción de valores, chocaríamos contra un nuevo problema: el problema de decidir qué valores introducir. En otras palabras, ¿qué queremos que quiera una superinteligencia? Este es el problema eminentemente filosófico al que nos dirigimos a continuación.

CAPÍTULO 13

Eligiendo los criterios para elegir

S

upongamos que pudiéramos instalar cualquier valor final arbitrario en una IA seminal. La decisión sobre qué valor instalar podría tener entonces consecuencias de lo más trascendental. Algunas otras opciones sobre los parámetros básicos —relativos a los axiomas de la teoría de la decisión de la IA y su epistemología — podrían tener consecuencias igualmente importantes. Pero con lo tontos, ignorantes y reduccionistas que somos, ¿cómo podemos confiar en que tomaremos buenas decisiones de diseño? ¿Cómo podríamos elegir sin fijar para siempre los prejuicios y las ideas preconcebidas de la generación actual? En este capítulo, exploramos cómo la normatividad indirecta puede permitirnos descargar gran parte del trabajo cognitivo implicado en la toma de estas decisiones en la propia superinteli- gencia, al tiempo que anclamos el resultado en los valores humanos más profundos.

La necesidad de normatividad indirecta

¿Cómo podemos hacer que una superinteligencia haga lo que queramos? ¿Qué queremos que quiera la superinteligencia? Hasta este punto, nos hemos centrado en la primera pregunta. Ahora abordaremos la segunda pregunta.

Supongamos que hubiéramos resuelto el problema del control de manera que fuéramos capaces de introducir cualquier valor que eligiéramos en el sistema de motivación de una superinteligencia para que persiguiera ese valor como su objetivo final. ¿Qué valor deberíamos introducir? La elección no es un tema baladí. Si la superinteligencia obtuviera una ventaja estratégica decisiva, el valor determinaría la disposición de los recursos cósmicos.

Claramente, es esencial que no nos equivoquemos en nuestra selección de valores. Pero ¿cómo podríamos, de manera realista, esperar alcanzar la infalibilidad en una cuestión como ésta? Podríamos estar equivocados acerca de la moralidad; equivocados también sobre lo que es bueno para nosotros; equivocados, incluso, sobre lo que realmente queremos. La especificación de un objetivo final, al parecer, requiere abrirse camino a través de una maraña de espinosos problemas filosóficos. Si intentamos una aproximación directa, probablemente nos haremos un lío. El riesgo de elegir equivocadamente es especialmente alto cuando el contexto de la decisión no es familiar —y la selección del objetivo final de una superinteligencia artificial que dará forma al futuro completo de la humanidad es un contexto de decisión extremadamente poco familiar, si alguno lo es.

Las deprimentes probabilidades de éxito en un asalto frontal se

reflejan en el disenso generalizado sobre los temas relevantes de la teoría del valor. No hay teoría ética que goce de apoyo mayoritario entre los filósofos, por lo que la mayoría de los filósofos deben estar equivocados.[1] También se refleja en los marcados cambios que la distribución de la creencia moral ha sufrido a lo largo del tiempo, muchos de los cuales nos gusta entenderlos como progreso. En la Europa medieval, por ejemplo, se consideró un entretenimiento respetable ver a un preso político ser torturado hasta la muerte. La quema de gatos siguió siendo popular en el París del siglo XVI.[2] Hace tan sólo ciento cincuenta años, la esclavitud todavía se practicaba ampliamente en el sur de Norteamérica, con el apoyo total de la ley y la costumbre moral. Cuando miramos hacia atrás, vemos deficiencias evidentes no sólo en el comportamiento, sino en las creencias morales de todas las edades anteriores. Aunque desde entonces quizá hayamos alcanzado cierta conciencia moral, difícilmente podríamos afirmar que estamos ahora en sol de mediodía de la iluminación moral perfecta. Muy probablemente, todavía estamos trabajando bajo uno o más conceptos morales gravemente erróneos. En tales circunstancias, seleccionar un valor final sobre la base de nuestras convicciones actuales, de manera que se fijara para siempre y excluyera cualquier posibilidad de mayor progreso ético, sería correr el riesgo de una calamidad moral existencial.

Incluso si pudiéramos estar racionalmente seguros de haber identificado la teoría ética correcta —algo que no es así— todavía estaríamos bajo el riesgo de cometer errores en la elaboración de importantes detalles de esta teoría. Teorías morales aparentemente simples pueden tener un montón de complejidad oculta.[3] Por ejemplo, consideremos la (especialmente simple) teoría consecuencialista del hedonismo. Esta teoría afirma, más o menos, que entre todas las cosas lo único que tiene valor es el placer, y que entre todas las cosas sólo el dolor tiene valor negativo.[4] Incluso si nos jugáramos todas fichas morales en esta teoría, y la teoría resultara ser correcta, un gran número de preguntas permanecerían abiertas. ¿Se debe dar a los "placeres superiores" prioridad sobre los "placeres inferiores", como argumentó John Stuart Mill? ¿Cómo debería tenerse en cuenta la intensidad y la duración de un placer? ¿Pueden los dolores y los placeres anularse entre sí? ¿Qué tipos de estados cerebrales se asocian con placeres moralmente relevantes? ¿Dos copias exactas del mismo estado cerebral corresponderían al doble de cantidad de placer?[5] ¿Puede haber placeres subconscientes? ¿Cómo debemos abordar ocasiones muy pequeñas de placeres muy grandes?[6] ¿Cómo deberíamos abordar poblaciones infinitas?[7]

Dar la respuesta equivocada a cualquiera de estas preguntas podría ser catastrófico. Si al seleccionar un valor final para la

superinteligencia tuviéramos que apostar, no sólo por una teoría moral general, sino por un extenso conjunto de reivindicaciones específicas sobre cómo esa teoría debe interpretarse e integrarse en un proceso de toma de decisiones efectivo, entonces nuestras posibilidades de salir ganadores empezarían a disminuir hasta algo cercano a la desesperanza. Los tontos podrían aceptar con entusiasmo este desafío de resolver de un golpe todos los problemas importantes de la filosofía moral, con el fin de fijar sus respuestas favoritas en la IA seminal. Almas más sabias trabajarían duro para encontrar algún enfoque alternativo, alguna manera de jugar sobre seguro.

Esto nos lleva a la normatividad indirecta. La razón obvia para construir una su- perinteligencia es para que podamos introducir en ella datos sobre la razón instrumental necesaria para encontrar formas eficaces de realizar un valor dado. La norma- tividad indirecta nos permitiría también descargar en la superinteligencia algunos de los razonamientos necesarios para seleccionar el valor que se quiere realizar.

La normatividad indirecta es una forma de responder al desafío presentado por el hecho de que quizá no sepamos lo que realmente queremos, lo que nos interesa, o lo que es moralmente correcto o ideal. En lugar de hacer una conjetura basada en nuestra propia comprensión actual (que es probablemente profundamente defectuosa), queremos delegar parte del trabajo cognitivo necesario para seleccionar valores a la superinteligencia. Dado que la superinteligencia es mejor que nosotros en el trabajo cognitivo, podría ver más allá de los errores y confusiones que nublan nuestro pensamiento. Podríamos generalizar esta idea y condensarla como principio heurístico:

> **El principio de deferencia epistémica**
> Una superinteligencia futura ocupa un punto de observación epistémicamente superior: sería de esperar que sus creencias fueran (probablemente, en la mayoría de los temas) más verdaderas que las nuestras. Deberíamos, por tanto, ceder ante la opinión de la superinteligencia cuando sea factible.[8]

La normatividad indirecta aplica este principio al problema de selección de valores. A falta de confianza en nuestra capacidad para especificar un estándar normativo concreto, especificaríamos en su lugar alguna condición más abstracta que cualquier estándar normativo satisfaría, con la esperanza de que una superinteligencia pudiera encontrar una norma concreta que satisficiera la condición abstracta. Podríamos dar a una IA seminal el objetivo final de actuar continuamente de acuerdo a su mejor estimación de lo que esta norma implícitamente definida le obligaría hacer.

Algunos ejemplos servirán para clarificar la idea. En primer lugar, vamos a considerar “la voluntad coherente extrapolada”, una

propuesta de normatividad indirecta esbozada por Eliézer Yudkowsky. Posteriormente, presentaremos algunas variaciones y alternativas, para darnos una idea de la gama de opciones disponibles.

Voluntad coherente extrapolada

Yudkowsky ha propuesto que se le diera a una IA seminal el objetivo final de llevar a cabo "la voluntad coherente extrapolada" de la humanidad (VCE), que se define de la siguiente manera:

> Nuestra voluntad coherente extrapolada serían nuestros deseos si supiéramos más, pensáramos más rápido, nos pareciéramos más a las personas que deseamos ser hubiéramos crecido más juntos; donde la extrapolación convergiera en lugar de divergir donde nuestros deseos cohesionaran en lugar de interferir; extrapolada como desearíamos que se extrapolara, interpretada como desearíamos que se interpretara.[9]

Cuando Yudkowsky escribió esto, no pretendía presentar un modelo para aplicar esta prescripción un tanto poética. Su objetivo era dar un bosquejo preliminar de cómo podría definirse la VCE, junto con algunos de los argumentos de por qué es necesario un enfoque de este tipo.

Muchas de las ideas detrás de la propuesta VCE tienen análogos y antecedentes en la literatura filosófica. Por ejemplo, en ética, las teorías del *observador ideal* tratan de analizar conceptos normativos como "bien" o "correcto" en relación a los juicios que un observador hipotético ideal haría (donde un "observador ideal» se define como uno que fuera omnisciente sobre hechos no-morales, fuera lógicamente clarividente, fuera imparcial de un modo relevante y estuviera libre de diversos tipos de sesgos, y así sucesivamente).[10] El enfoque de la VCE, sin embargo, no es (o no necesita ser interpretado como) una teoría moral. No está comprometida con la afirmación de que exista algún vínculo necesario entre el valor y las preferencias de nuestra voluntad coherente extrapolada. La VCE puede considerarse simplemente como una forma útil de aproximarse a cualquier cosa que tenga valor final, o puede considerarse como algo totalmente desconectado de la ética. Como prototipo principal del enfoque de normatividad indirecta, vale la pena examinarla con un poco más de detalle.

Algunas explicaciones

Algunos términos de la cita anterior requieren explicación. "Pensar más rápido", en la terminología de Yudkowsky significa *si fuéramos más inteligentes y meditáramos* más las cosas. "Crecer más juntos"

parece significar *si hubiéramos realizado nuestro aprendizaje, nuestro potenciamiento cognitivo, y nuestra auto-mejora en condiciones de adecuada interacción social con los demás.*

"Donde la extrapolación converge más que diverge" puede entenderse de la siguiente manera. La IA debería actuar de acuerdo al resultado de su extrapolación sólo en la medida en que la función pueda ser predicha por la IA con un grado bastante alto de confianza. En la medida en que la IA no pueda predecir lo que desearíamos si fuéramos idealizados de la manera indicada, la IA no debería actuar en base a ninguna conjetura salvaje; en su lugar, debería abstenerse de actuar. Sin embargo, a pesar de que muchos detalles de nuestros deseos idealizados pudieran ser indeterminados o impredecibles, podría, no obstante, haber algunos grandes rasgos que la IA podría aprehender, y podría por lo menos actuar para garantizar que el curso futuro de los acontecimientos se desarrollara dentro de esos parámetros. Por ejemplo, si la IA pudiera estimar con fiabilidad que nuestra voluntad coherente extrapolada desearía que no estuviéramos todos en agonía constante, o que el universo no fuera alicatado con clips, entonces la IA debería actuar para evitar esos resultados.[11]

"Cuando nuestros deseos cohesionaran en lugar de interferir" puede interpretarse de la siguiente manera. La IA debe actuar donde haya un amplio acuerdo entre voluntades extrapoladas de seres humanos individuales. Un conjunto más pequeño de deseos fuertes y claros podrían a veces imponerse a los deseos débiles y confusos de una mayoría. Asimismo, Yudkowsky piensa que debería requerir menos consenso que la IA *previniera* algún resultado particularmente especificado, y más consenso para que la IA actuara canalizando el futuro en base a alguna estrecha concepción del bien". "La dinámica inicial de la VCE", escribe, "debe ser reticente a decir «sí», y escuchar con atención para decir «no»".[12]

"Extrapolada como desearíamos que se extrapolara, interpretada como desearíamos que se interpretara": La idea detrás de estos últimos modificadores parece ser que las reglas para la extrapolación deben ser ellas mismas sensibles a la voluntad extrapolada. Un individuo puede tener un deseo de segundo orden (un deseo respecto a qué desear) de que parte de sus deseos de primer orden no fueran tenidos en cuenta cuando se extrapolara su voluntad. Por ejemplo, un alcohólico que tuviera un deseo de primer orden por el alcohol, también podría tener un deseo segundo orden de no tener ese deseo de primer orden. Del mismo modo, podríamos tener deseos sobre cómo deben desarrollarse algunas otras partes del proceso de extrapolación, y éstos deberían tenerse en cuenta en el proceso de extrapolación.

Se podría objetar que, aunque el concepto de voluntad coherente

extrapolada de la humanidad pudiera definirse correctamente, de todos modos sería imposible —incluso para una superinteligencia— averiguar lo que la humanidad querría realmente en las circunstancias hipotéticas ideales estipuladas en el enfoque de la VCE. Sin alguna información sobre el contenido de nuestra voluntad extrapolada, la IA carecería de normas relevantes que guiaran su comportamiento. Sin embargo, aunque fuera difícil saber con precisión lo que desearía la VCE de la humanidad, es posible hacer conjeturas informadas. Esto es posible incluso hoy en día, sin superinteligencia. Por ejemplo, es más plausible que nuestra VCE deseara que hubiera gente en el futuro que viviera vidas ricas y felices, a que deseara que todos estuviéramos sentados en los taburetes de una habitación oscura experimentando dolor. Si somos capaces de hacer por lo menos algunos juicios sensatos de este tipo, también podría una superinteligencia. Desde el principio, la conducta de la superinteligencia podría por lo tanto ser guiada por sus estimaciones sobre el contenido de nuestra VCE. Tendría fuertes razones instrumentales para refinar estas estimaciones iniciales (por ejemplo, mediante el estudio de la cultura humana y la psicología, el escaneo de cerebros humanos, y el razonamiento acerca de cómo podríamos comportarnos si supiéramos más, pensáramos con más claridad, etc.). En la investigación de estos asuntos, la IA se guiaría por sus estimaciones iniciales sobre nuestra VCE; de modo que, por ejemplo, la IA no ejecutaría innecesariamente innumerables simulaciones repletas de sufrimiento humano irredento si estimara que nuestra VCE probablemente condenaría tales simulaciones como crímenes mentales.

Otra objeción es que hay muchas maneras diferentes de vida y de códigos morales en el mundo que podrían ser imposibles de “mezclar” en una sola VCE. Incluso si se pudiera mezclarlos, el resultado podría no ser particularmente apetecible —sería poco probable conseguir una deliciosa comida mezclando todos los mejores sabores de los platos favoritos de todo el mundo.[13] En respuesta a esto, se podría señalar que el enfoque de VCE no requiere que todas las formas de vida, códigos morales o valores personales sean mezclados juntos en un guiso. La dinámica de la VCE implica que sólo debería actuar cuando nuestros deseos fueran coherentes. En cuestiones en las que existiera un desacuerdo irreconciliable generalizado, incluso después de que se hubieran impuesto diversas condiciones idealizadoras, la dinámica debería abstenerse de determinar el resultado. Para continuar con la analogía de la cocina, podría ser que las personas o las culturas tengan diferentes platos favoritos, pero que, sin embargo, en términos generales, se pongan de acuerdo sobre qué alimentos deberían ser considerados no tóxicos. La dinámica de la VCE podría entonces actuar para prevenir la intoxicación alimentaria, mientras que permite

que los seres humanos resuelvan de otra manera sus prácticas culinarias sin su orientación o interferencia.

Razones para la VCE

El artículo de Yudkowsky ofreció siete argumentos a favor del enfoque VCE. Tres de ellos eran básicamente diferentes maneras de expresar el punto de que, si bien el objetivo debería consistir en hacer algo que fuera humano y servicial, sería muy difícil establecer un conjunto explícito de reglas que no tuvieran interpretaciones involuntarias y consecuencias indeseables.[14] El enfoque VCE está pensado para ser robusto y corregirse a sí mismo; está pensado para captar la *fuente* de nuestros valores en lugar de confiar en que nosotros enumeremos y articulemos correctamente, de una vez por todas, cada uno de nuestros valores esenciales.

Los cuatro restantes argumentos van más allá de ese primer (pero importante) punto básico, exponiendo desiderata sobre las posibles soluciones al problema de especificación de valor y sugiriendo que la VCE cumpliría con estos desiderata.

"Preservar el crecimiento moral"

Éste es el desiderátum de que la solución debería hacer posible el progreso moral. Como se sugirió anteriormente, hay razones para creer que nuestras creencias morales actuales están viciadas de muchas maneras; quizá profundamente viciadas. Si tuviéramos que estipular un código moral concreto e inalterable para la IA, estaríamos fijando nuestras actuales convicciones morales, incluyendo sus errores, destruyendo cualquier esperanza de crecimiento moral. El enfoque de la VCE, por el contrario, hace posible tal crecimiento en tanto que la IA trataría de hacer lo que nos hubiera gustado hacer si nos hubiéramos desarrollado en condiciones aún más favorables, y es posible que si hubiéramos desarrollado nuestras creencias y sensibilidades morales, habrían sido purgadas de sus defectos y limitaciones actuales.

"Evitar que se secuestre el destino de la humanidad"

Yudkowsky tiene en mente un escenario en el que un pequeño grupo de programadores creara una IA seminal que luego se convirtiera en una superinteligencia que obtuviera una ventaja estratégica decisiva. En este escenario, los programadores originales tendrían en sus manos la totalidad de los recursos cósmicos de la humanidad. Obviamente, ésta es una responsabilidad horrible con la que cargar a cualquier mortal. Sin embargo, no es posible que los programadores eludan por completo la responsabilidad una vez se encuentren en esa situación:

cualquier elección que hicieran, incluyendo el abandono del proyecto, tendría consecuencias históricas mundiales. Yudkowsky ve la VCE como una manera de que los programadores eviten arrogarse para sí el privilegio o la responsabilidad de determinar el futuro de la humanidad. Con la creación de una dinámica que implementara la voluntad coherente extrapolada de la *humanidad* —en oposición a implementar su propia voluntad, o su teoría moral favorita —se distribuye, en efecto, la influencia sobre el futuro a toda la humanidad.

"Evitar la creación de motivos para que los seres humanos de hoy en día se peleen por la dinámica inicial"

La distribución de influencia sobre el futuro de la humanidad no sólo es moralmente preferible a que el equipo de programación implemente su propia visión favorita, también es una forma de reducir el incentivo de luchar por ser quién cree la primera superinteligencia. En el enfoque de la VCE, los programadores (o sus patrocinadores) no ejercen más influencia sobre el contenido de los resultados que cualquier otra persona, a pesar de que, por supuesto, desempeñan un papel causal protagonista en la determinación de la estructura de extrapolación y en la decisión de aplicar la VCE de la humanidad en lugar de alguna alternativa. Evitar el conflicto es importante no sólo por el daño inmediato que el conflicto suele provocar, sino también porque obstaculiza la colaboración en el difícil reto de desarrollar una superinteligencia segura y beneficiosa.

La VCE está pensada para contar con un amplio apoyo. Esto no es sólo porque asigne una influencia equitativa. También hay una base más profunda de potencial conciliador de la VCE, a saber, que permite a muchos grupos diferentes a la esperanza de que su visión preferida del futuro prevalecerá totalmente. Imaginemos un miembro de los talibanes afganos debatiendo con un miembro de la Asociación Sueca Humanista. Los dos tienen visiones del mundo muy diferentes, y lo que es una utopía para uno podría ser una distopía para el otro. Podría ser que tampoco les entusiasmara ninguna posición de compromiso, tales como permitir a las niñas recibir una educación, pero sólo hasta el noveno curso, o permitir que hubiera educación para las niñas suecas, pero no para las niñas afganas. Sin embargo, tanto el talibán como el humanista podrían ser capaces de respaldar el principio de que el futuro debe ser determinado por la VCE de la humanidad. Los talibanes podría razonar que si sus puntos de vista religiosos son de hecho correctos (de lo que están convencidos) y si existen buenas razones para aceptar estos puntos de vista (de lo que también están convencidos), entonces la humanidad llegaría finalmente a aceptar estos puntos de vista, si la gente tuviera menos

prejuicios y sesgos, si pasaran más tiempo estudiando las escrituras, si pudieran comprender más claramente cómo funciona el mundo y reconocer las prioridades esenciales, si pudieran ser liberados de la rebeldía irracional y la cobardía, etc.[15] El humanista, de manera similar, creería que en estas condiciones idealizadas, la humanidad habría llegado a aceptar los principios que propugna.

"Mantener a la humanidad a cargo de su propio destino en última instancia"

Puede que no queramos un resultado en el que una superinteligencia paternalista nos vigile constantemente, microgestionando nuestros asuntos con el único objetivo de optimizar cada detalle de acuerdo con un gran plan. Incluso si estipuláramos que la superinteligencia fuera a ser perfectamente benevolente y libre de presunción, arrogancia, prepotencia, estrechez de miras, y otros defectos humanos, todavía se podría resentir la pérdida de autonomía que conlleva tal disposición. Podríamos preferir crear nuestro destino a medida que avanzamos, incluso si esto significa que a veces tropecemos. Quizás querríamos que la superinteligencia sirviera como una red de seguridad, para que nos apoyara cuando las cosas fueran catastróficamente mal, pero que por lo demás nos dejara manejarnos por nosotros mismos.

La VCE permite esta posibilidad. La VCE está destinada a ser una "base dinámica", un proceso que se ejecuta una vez y luego se reemplaza con lo que la voluntad extrapolada desea. Si la voluntad extrapolada de la humanidad deseara que viviéramos bajo la supervisión de una IA paternalista, entonces la dinámica VCE crearía una IA así y le entregaría las riendas. Si la voluntad extrapolada de la humanidad en su lugar deseara que se creara un gobierno mundial democrático humano, entonces la dinámica VCE podría facilitar el establecimiento de una institución de este tipo y ser invisible para el resto de cuestiones. Si la voluntad extrapolada de la humanidad fuera, en cambio, que cada persona debiera recibir una dotación de recursos que pudiera usar de la manera que quisiera, siempre y cuando respetara la igualdad de derechos de los demás, entonces la dinámica VCE podría hacer que esto se hiciera realidad al operar en el trasfondo, de manera muy similar a una ley de la naturaleza, para evitar la entrada ilegal, el robo, el asalto, y otras transgresiones no consensuadas.[16]

La estructura del enfoque VCE permitiría, por tanto, una gama prácticamente ilimitada de resultados. También es concebible que la voluntad extrapolada de la humanidad deseara que la VCE no hiciera nada en absoluto. En ese caso, la IA implementa- dora de la VCE debería, al haber establecido con suficiente probabilidad que esto es lo que la voluntad extrapolada de la humanidad desearía hacer, apagarla

con seguridad.

Otras observaciones

La propuesta de la VCE, como se indicó anteriormente, es, por supuesto, muy esquemática. Tiene un número de parámetros libres que podrían especificarse de varias formas, produciendo versiones diferentes de la propuesta.

Un parámetro es la base de extrapolación: ¿Las voluntades de quiénes serán incluidas? Podríamos decir "todo el mundo", pero esta respuesta engendra una serie de preguntas adicionales. ¿La base de extrapolación incluye las llamadas "personas marginales" como embriones, fetos, personas con muerte cerebral, los pacientes con demencias graves o que se encuentran en estados vegetativos permanentes? ¿Tiene cada uno de los hemisferios de un paciente con "cerebro dividido" su propio peso en la extrapolación y es este peso el mismo que el de todo el cerebro de un sujeto normal? ¿Qué pasa con las personas que vivieron en el pasado pero ahora están muertas? ¿Con las personas que nacerán en el futuro? ¿Con los animales superiores y otras criaturas sensibles? ¿Con las mentes digitales? ¿Con los extraterrestres?

Una opción sería incluir sólo a la población de seres humanos adultos de la Tierra que están vivos al inicio de la época en que se cree la IA. Una extrapolación inicial de esta base podría entonces decidir si la base debe ser ampliada y cómo. Dado que el número de "marginales" en la periferia de esta base es relativamente pequeño, el resultado de la extrapolación no puede depender mucho de donde se dibuje exactamente el límite —de si, por ejemplo, incluye o no a fetos.

Que alguien sea excluido de la base original de extrapolación no implica que sus deseos y su bienestar no sean tenidos en cuenta. Si la voluntad coherente extrapolada de los que están en la base de extrapolación (por ejemplo, los seres humanos adultos vivientes) desearan que la consideración moral se extendiera a otros seres, entonces el resultado de la dinámica VCE reflejaría esa preferencia. Sin embargo, es posible que los intereses de los que están incluidos en la base de extrapolación original sean tenidos en cuenta en un grado mayor que los intereses de los forasteros. En particular, si la dinámica actuara sólo cuando existiera un amplio acuerdo entre voluntades individuales extrapoladas (como en la propuesta original de Yudkowsky), parece que habría un riesgo significativo de un voto de bloqueo poco generoso que podría evitar, por ejemplo, el bienestar de los animales no humanos o que las mentes digitales fueran protegidas. El resultado podría estar moralmente podrido.[17]

Una de las motivaciones para la propuesta VCE era evitar la

creación de un motivo para que los humanos se pelearan por ser los primeros en crear una IA superinteligen- te. Aunque la propuesta VCE parezca mejor respecto de este desiderátum que muchas alternativas, no elimina por completo los motivos de conflicto. Un individuo, grupo o nación egoísta podría tratar de ampliar su parte del futuro, dejando a otros fuera de la base de extrapolación.

Una toma de poder de este tipo podría ser racionalizada de varias maneras. Se podría argumentar, por ejemplo, que el patrocinador que financia el desarrollo de la
IA merece ser dueño del resultado. Esta reivindicación moral es probablemente falsa. Podría objetarse, por ejemplo, que el proyecto que pusiera en marcha la primera IA seminal exitosa impondría un gran riesgo externo al resto de la humanidad, que, por tanto, tiene derecho a una indemnización. El monto de la indemnización adeudada sería tan grande que sólo podría compensarse dando a todos una participación del proyecto si todo saliera bien.[18]

Otro argumento que podría utilizarse para racionalizar la toma de poder es que grandes segmentos de la humanidad tienen una base maligna o preferencias por el mal y que su inclusión en la base de extrapolación introduciría un riesgo de que el futuro de la humanidad se convirtiera en una distopía. Es difícil saber la cuota del bien y del mal en el corazón de la persona promedio. También es difícil saber cómo varía este equilibrio entre los diferentes grupos, estratos sociales, culturas o naciones. Dependiendo de si se es optimista o pesimista sobre la naturaleza humana, se puede preferir no apostar los recursos cósmicos de la humanidad especulando que, en una mayoría suficiente de los siete mil millones de personas que actualmente viven, prevalecerá su parte buena al extrapolar su voluntad. Por supuesto, la omisión de un determinado conjunto de personas de la base de extrapolación no garantiza que la luz triunfara; y bien podría ser que las almas que excluyeran a otros o que tomaran el poder para sí mismos tiendan a tener cantidades inusualmente grandes de oscuridad.

Sin embargo, otra razón para luchar por la dinámica base es que uno podría creer que la IA de otro no funciona como se cree, incluso si la IA fuera considerada como una forma de poner en práctica la VCE de la humanidad. Si los diferentes grupos tuvieran diferentes creencias acerca de qué aplicación es más probable que tenga éxito, podrían luchar para evitar que los otros la lanzaran. En este tipo de situaciones, sería mejor que los proyectos en competencia pudieran resolver sus diferencias epistémicas por algún método que determinara —de manera más fiable que mediante el conflicto armado — qué es justo.[19]

Modelos de moralidad

La propuesta VCE no es la única forma posible de normatividad indirecta. Por ejemplo, en lugar de aplicar la voluntad coherente extrapolada de la humanidad, se podría tratar de construir una IA con el objetivo de hacer lo que fuera moralmente correcto, basándose en las capacidades cognitivas superiores de la IA para averiguar exactamente qué acciones se ajustan a esa descripción. Podemos llamar a esta propuesta "rectitud moral" (RM). La idea es que los seres humanos tienen una comprensión imperfecta de lo que es correcto y lo incorrecto, y tal vez una comprensión aún más pobre de cómo se analiza filosóficamente el concepto de rectitud moral: pero una superinteligencia podría comprender mejor estas cosas.[20]

¿Y si no estamos seguros de que el realismo moral sea cierto? Todavía podríamos intentar la propuesta RM. Sólo deberíamos asegurarnos de especificar lo que la IA debería hacer en la eventualidad de que su presupuesto de realismo moral fuera falso. Por ejemplo, podríamos establecer que si la IA calculara con una probabilidad suficiente que no hay verdades no-relativas sobre la rectitud moral, entonces debería volver a la implementación de la voluntad coherente extrapolada en su lugar, o simplemente apagarse.[21]

La RM parece tener varias ventajas sobre la VCE. La RM acabaría con diversos parámetros libres en la VCE, como el grado de coherencia entre las voluntades extrapoladas que se requerirían para que la IA actuara sobre el resultado, la facilidad con que una mayoría podría anular a las minorías disidentes, y la naturaleza del entorno social dentro del cual se supone que nuestros yoes extrapolados deberían haber "crecido más juntos" Al parecer, eliminaría la posibilidad de un fracaso moral resultante de la utilización de una base de extrapolación demasiado estrecha o demasiado ancha. Además, la RM orientaría a la IA hacia la acción moralmente correcta incluso si nuestras voluntades coherentes extrapoladas desearan que la IA tomara acciones que fueran moralmente odiosas. Como se señaló anteriormente, esto parece una posibilidad factible con la propuesta de la VCE. La bondad moral podría ser más un metal precioso que un elemento abundante en la naturaleza humana, e incluso después de que el mineral hubiera sido procesado y refinado de acuerdo con las prescripciones de la propuesta VCE, ¿quién sabe si el resultado principal será una brillante virtud, escoria indiferente, o lodos tóxicos?

La RM también parece tener algunas desventajas. Se basa en la noción de lo "moralmente correcto", un concepto muy difícil, con el que los filósofos han lidiado desde la antigüedad sin haber alcanzado todavía un consenso en sus análisis. Tomar una explicación errónea de

"rectitud moral" podría dar lugar a resultados que serían moralmente muy negativos. Esta dificultad de definir "rectitud moral" parece que podría contar muy en contra de la propuesta RM. Sin embargo, no está claro que la propuesta RM esté realmente en desventaja material en este sentido. La propuesta de la VCE, también utiliza términos y conceptos que son difíciles de explicar (como "conocimiento", "ser más las personas que nos hubiera gustado ser", "crecer más juntos", entre otros).[22] Aunque estos conceptos son marginalmente menos opacos que "rectitud moral", no dejan de estar muy alejados de cualquier cosa que los programadores puedan expresar actualmente en código.[23] El camino para dotar a una IA con alguno de estos conceptos podrían implicar darle capacidad lingüística general (comparable, por lo menos, a la de un ser humano adulto normal).Tal capacidad general para entender el lenguaje natural podría utilizarse para comprender lo que se entiende por "moralmente correcto". Si la IA pudiera captar el significado, podría buscar acciones que se ajustaran. A medida que la IA desarrollara la superinteligencia, podría entonces avanzar en dos frentes: en el problema filosófico de comprensión de qué es la rectitud moral, y en el problema práctico de aplicar este conocimiento para evaluar acciones en particular.[24] Aunque esto no vaya a ser fácil, no está claro que fuera *más* difícil que extrapolar la voluntad coherente extrapolada de la humanidad.[25]

Una cuestión más fundamental con la RM es que incluso si se pudiera implemen- tar, puede que no nos diera lo que queremos o lo que elegiríamos si fuéramos más brillantes y estuviéramos mejor informados. Esto es, por supuesto, la característica esencial de la RM, no un error accidental. Sin embargo, podría ser una característica que acabara siendo extremadamente perjudicial para nosotros.[26]

Se podría tratar de preservar la idea básica del modelo RM reduciendo a la vez su exigencia, centrándonos en la *permisibilidad moral:* con la idea de que podríamos dejar que la IA persiguiera la VCE de la humanidad con tal de que no actuara en formas que fueran moralmente inaceptables. Por ejemplo, se podría formular el siguiente objetivo para la IA:

Entre las acciones que son moralmente permisibles para la IA, elige la que VCE de la humanidad preferiría. Sin embargo, si una parte de esta instrucción no tuviera un significado bien especificado, o si estuviéramos radicalmente confundidos acerca de su significado, o si el realismo moral fuera falso, o si actuamos de una manera moralmente inadmisible en la creación de una IA con este objetivo, sométase entonces a un apagado controlado.[27] Siga el significado pretendido de esta instrucción.

Uno podría todavía preocuparse de que este modelo de permisibilidad moral (PM) representara un grado intransigentemente alto de respeto a las exigencias morales. La magnitud del sacrificio que sería necesario dependería de qué teoría ética fuera la verdadera.[28] Si la ética fuera

satisfaciente, en el sentido de que contara como moralmente permisible cualquier acción que se ajustara a unas pocas restricciones morales básicas, entonces la PM podría dejar un amplio espacio para que nuestra voluntad coherente extrapolada influyera en las acciones de la IA. Sin embargo, si la ética fuera *maximizante* —por ejemplo, si las únicas acciones moralmente admisibles fueran las que tuvieran mejores consecuencias morales— entonces, la PM podría dejar poco o ningún espacio para que nuestras propias preferencias conformaran el resultado.

Para ilustrar este problema, volvamos por un momento al ejemplo del consecuen- cialismo hedonista. Supongamos que esta teoría ética es cierta, y que la IA sabe que es así. A los efectos presentes, podemos definir el consecuencialismo hedonista como la afirmación de que una acción es moralmente correcta (y moralmente permisible) si y sólo si, entre todas las acciones posibles, ninguna otra acción produjera una mayor preponderancia del placer sobre el sufrimiento. La IA, siguiendo la PM, puede maxi- mizar el exceso de placer convirtiendo el universo accesible en hedonium, un proceso que puede implicar la construcción de computronium y utilizarlo para realizar cálculos que hagan las veces de experiencias placenteras. Ya que la simulación de todos los cerebros humanos existentes no es la manera más eficiente de producir placer, una consecuencia probable es que todos muramos.

Al promover la propuesta RM o la PM, correríamos por lo tanto el riesgo de sacrificar nuestras vidas por un bien mayor. Esto sería un sacrificio más grande de lo que se podría pensar, porque lo que nos arriesgamos a perder no es simplemente la oportunidad de vivir una vida humana normal, sino la oportunidad de disfrutar de las lejanas y más ricas vidas que una superinteligencia amigable podría otorgarnos.

El sacrificio parece aún menos atractivo cuando nos percatamos de que la super- inteligencia podría realizar un bien casi idéntico (en términos fraccionarios), sacrificando mucho menos de nuestro potencial bienestar. Supongamos que nos pusimos de acuerdo en permitir que *casi* todo el universo accesible fuera convertido en hedonium —todo, excepto una pequeña reserva, digamos la Vía Láctea, que se reservaría para dar cabida a nuestras propias necesidades. Entonces todavía habría cien mil millones de galaxias dedicadas a la maximización de placer. Pero tendríamos una galaxia en la que crear civilizaciones maravillosas que podrían durar por miles de millones de años y en las que los seres humanos y los animales no humanos podríamos sobrevivir y prosperar, y tener la oportunidad de convertirnos en santos espíritus posthumanos.[29]

Si alguno prefiriera esta última opción (como yo me inclinaría a hacer) ello implicaría que no tenemos una preferencia léxicamente dominante e incondicional para actuar de manera moralmente

permisible. Pero sería consistente con dar un gran peso a la moralidad.

Incluso desde un punto de vista puramente moral, tal vez sería mejor defender alguna propuesta que fuera menos moralmente ambiciosa que la RM o la PM. Si la moralmente mejor no tiene ninguna posibilidad de ser implementada —tal vez debido a sus exigencias antipáticas— podría ser moralmente mejor *promover* alguna otra propuesta, que fuera casi ideal y cuyas posibilidades de ser implementada podrían aumentar significativamente gracias a nuestro apoyo.[30]

Haz lo que quiero que hagas

Podemos sentirnos inseguros de apostar por la VCE, la RM o la PM, o alguna otra. ¿Podríamos desentendernos también de esta decisión de nivel superior, descargando aún más trabajo cognitivo en la IA? ¿Dónde está el límite de nuestra posible pereza?

Considérese, por ejemplo, el siguiente objetivo "basado en razones":

Haz aquello que hubiera sido más razonable que le pidiéramos hacer a la IA.

Este objetivo podría acabar llevándonos a la voluntad extrapolada o a la moral o a alguna otra cosa, pero parece que nos ahorraría el esfuerzo y el riesgo involucrado en tratar de averiguar por nosotros mismos cuál de estos objetivos específicos hubiéramos tenido más razones para seleccionar.

Algunos de los problemas de los objetivos basados en la moralidad, sin embargo, también se aplican aquí. En primer lugar, podríamos temer que este objetivo basado en razones dejara muy poco espacio para nuestros propios deseos. Algunos filósofos sostienen que una persona siempre tiene razones para hacer lo que sería moralmente mejor para ella. Si esos filósofos tuvieran razón, entonces el objetivo —basado en razones colapsaría en RM— con el consiguiente riesgo de que una superinteligencia implementara una dinámica tal que matara a todos a su alcance. En segundo lugar, al igual que con todas las propuestas formuladas en lenguaje técnico, hay una posibilidad de que pudiéramos haber entendido mal el sentido de nuestras propias afirmaciones. Hemos visto que, en el caso de los objetivos basados en la moral, pedirle a la IA que haga lo correcto puede acarrear consecuencias imprevistas e indeseadas de tal manera que, si las hubiéramos previsto, no hubiéramos implementado el objetivo en cuestión. Lo mismo se aplicaría a pedirle a la IA que hiciera lo que tuviéramos más razones para hacer.

Qué pasaría si tratáramos de evitar estas dificultades proponiendo

un objetivo en lenguaje enfáticamente no técnico —del tipo "agradabilidad": [31]

> Lleva a cabo la acción más agradable; o, si ninguna acción es la más agradable, lleva a cabo entonces una acción que sea al menos súper-mega agradable.

¿Cómo podría haber nada objetable a construir una IA agradable? Pero debemos preguntarnos qué se entiende exactamente con esta expresión. El léxico enumera diversos significados de "agradable" que claramente no están destinados a ser utilizados aquí: no tenemos la intención de que la IA sea *cortés y educada, ni delicada, ni fastidiosa.* Si pudiéramos contar con que la IA reconocería la interpretación pretendida de "amabilidad" y estuviera motivada para perseguir la amabilidad sólo en ese sentido, entonces este objetivo parecería equivaler a una orden para hacer lo que los programadores de la IA quisieran que hiciera.[32] Una orden de efecto similar se incluía en la formulación de la VCE ("... interpretada como desearíamos que se interpretara") y en el criterio de permisibilidad moral, como expusimos antes ("... seguir el significado pretendido de esta instrucción"). Al colocar dicha cláusula "haz lo que quiero que hagas" podemos indicar que las otras palabras en la descripción de objetivos se deben interpretar generosamente en lugar de literalmente. Pero decir que la IA debería ser "agradable", no añade casi nada: el verdadero trabajo se hace mediante la instrucción "haz lo que quiero que hagas". Si supiéramos cómo codificar "haz lo que quiero que hagas" de una manera general y de gran alcance, podríamos usar eso perfectamente como un único objetivo.

¿Cómo podría implementarse una dinámica de "haz lo que quiero que hagas"? Es decir, ¿cómo podríamos crear una IA motivada para interpretar caritativamente nuestros deseos e intenciones tácitas y actuar en consecuencia? Un paso inicial podría ser tratar de obtener más claridad acerca de lo que entendemos por "haz lo que quiero que hagas". Podría ayudar si pudiéramos explicar esto en términos más conductistas, por ejemplo, en términos de preferencias reveladas en varias situaciones hipotéticas —como situaciones en las que tuvimos más tiempo de considerar las opciones, en las que fuimos más inteligentes, en las que sabíamos más sobre los hechos pertinentes, y en las que, de varias distintas maneras, las condiciones fueron más favorables para que nosotros manifestáramos con precisión y de manera concreta lo que quisimos decir cuando dijimos que queríamos que una IA fuera amigable, benéfica, agradable...

Aquí, por supuesto, volvemos al punto de partida. Hemos vuelto a la aproximación de la normatividad indirecta con la que empezamos, la propuesta de la VCE, que, en esencia, elimina todo el contenido concreto de la especificación de valor, dejando sólo un valor abstracto

definido en términos puramente procedimentales: hacer lo que hubiéramos deseado que la IA hiciera en circunstancias debidamente idealizadas. Por medio de tal normatividad indirecta, podríamos esperar poder descargar en la IA gran parte del trabajo cognitivo que nosotros mismos estaríamos tratando de realizar si intentáramos articular una descripción más concreta de qué valores debería perseguir una IA. Al tratar de sacar el máximo provecho de la superioridad epistémica de la IA, la VCE puede, por lo tanto, ser vista como una aplicación del principio de deferencia epistémica.

Lista de componentes

Hasta ahora hemos considerado diferentes opciones sobre qué contenidos dar al sistema de objetivos. Pero el comportamiento de una IA también se vería influenciado por otras opciones de diseño. En particular, qué teoría de la decisión y qué epistemología utilizara podría marcar una diferencia crítica. Otra cuestión importante es si los planes de la IA serían revisados por humanos antes de llevarse a cabo.

La tabla 13 resume estas opciones de diseño. Un proyecto que tuviera como objetivo construir una superinteligencia debería ser capaz de explicar qué decisiones se tomaron sobre cada uno de estos componentes, y justificar por qué se tomaron esas decisiones.[33]

Tabla 13. ***Lista de componentes***

Contenido de los objetivos	¿Qué objetivos debería perseguir la IA? ¿Cómo deberían interpretarse las descripciones de los objetivos? ¿Debería incluirse entre los objetivos premiar especialmente a aquellos que contribuyan al éxito del proyecto?
Teoría de la decisión	¿Debería la IA utilizar la teoría causal de decisión, la teoría de la decisión evidencial, la teoría de la decisión no-actualizable, o alguna otra?
Epistemología	¿Qué función de probabilidad de principios debería tener la IA, y qué otras suposiciones, explícitas o implícitas, sobre el mundo debería hacer? ¿Qué teoría antrópica debería usar?
Deferencia	¿Deberían los planes de la IA ser sometidos a revisión humana antes de ser puestos en práctica? Si es así, ¿cuál es el protocolo para ese proceso de revisión?

Contenido de los objetivos

Ya hemos hablado de cómo la normatividad indirecta podría utilizarse en la especificación de valores que la IA debería perseguir. Discutimos algunas opciones, como los modelos basados en la moralidad y en la voluntad coherente extrapolada. Cada una de estas decisiones crea nuevas decisiones que necesitan ser afrontadas. Por ejemplo, el enfoque VCE se plasma de muchas formas, dependiendo de quién esté incluido en la base de extrapolación, la estructura de extrapolación, etc. Otros métodos de selección de la motivación podrían conllevar

diferentes tipos de contenido en los objetivos. Por ejemplo, un oráculo podría ser construido para que otorgara valor a dar respuestas precisas. Un oráculo construido con una motivación de domesticidad también podría tener un contenido de objetivos que valorara negativamente el uso excesivo de recursos en la producción de respuestas.

Otra opción de diseño consistiría en dilucidar si se deberían incluir disposiciones especiales en el contenido de los objetivos para premiar a los individuos que contribuyeran a la realización exitosa de la IA, dándoles, por ejemplo, recursos o influencia adicional sobre el comportamiento de la IA. Podemos denominar dichas disposiciones como "entrelazamiento de incentivos". El entrelazamiento de incentivos podría ser visto como una forma de aumentar la probabilidad de que el proyecto tenga éxito, a costa de comprometer en cierta medida el objetivo que el proyecto se propuso alcanzar.

Por ejemplo, si el objetivo del proyecto fuera crear una dinámica que implemen- tara la voluntad coherente extrapolada de la humanidad, entonces un esquema de entrelazamiento de incentivos podría especificar que las voluntades de ciertos individuos deberían tener un peso extra en la extrapolación. Si este proyecto tuviera éxito, el resultado no sería necesariamente la aplicación de la voluntad coherente extrapolada de la humanidad. En su lugar, alguna aproximación a este objetivo podría lograrse.[34]

Puesto que el entrelazamiento de incentivos sería una parte del contenido de los objetivos que sería interpretada y buscada por una superinteligencia, podría tomar ventaja de la normatividad indirecta para especificar disposiciones sutiles y complicadas que serían difíciles de implementar para un agente humano. Por ejemplo, en lugar de recompensar a los programadores según alguna métrica cruda pero fácilmente accesible, como el número de horas que trabajaran o cuántos errores se corrigieran, el entrelazamiento de objetivos podría especificar que los programadores "deberían ser recompensados en proporción a lo que sus contribuciones ayudaron, según cierta probabilidad *ex ante* razonable, a que el proyecto se completara con éxito de la manera que los patrocinadores querían" Además, no habría ninguna razón para limitar el entrelazamiento de incentivos al personal del proyecto. En su lugar, podría especificarse que cada persona debería ser recompensada según sus méritos. La asignación de crédito es un problema difícil, pero podríamos esperar que una superinteligencia hiciera un trabajo razonable aproximándose a los criterios especificados, explícita o implícitamente, por el entrelazamiento de incentivos.

Es concebible que la superinteligencia pudiera incluso encontrar alguna manera de recompensar a las personas que hubieran muerto

antes de la creación de la superinte- ligencia.[35] El entrelazamiento de incentivos podría entonces ampliarse para abarcar al menos a algunos de los fallecidos, potencialmente incluyendo individuos que murieron antes de que el proyecto fuera concebido, o incluso anteriores a la primera enunciación del concepto de entrelazamiento de incentivos. Aunque la institución de tal política retroactiva no fuera a incentivar causalmente a aquellas personas que ya estén descansando en sus tumbas mientras estas palabras se están escribiendo, podría ser favorecida por razones morales —aunque se podría argumentar que en la medida en que la equidad sea un objetivo, debería ser incluida como parte de la propia especificación de objetivo en lugar de incluirla en el entrelazamiento de incentivos circundante.

No podemos ahondar aquí en todas las cuestiones éticas y estratégicas relacionadas con el entrelazamiento de incentivos. El posicionamiento de un proyecto respecto de estos temas, sin embargo, sería un importante aspecto de su concepto fundamental de diseño.

Teoría de la decisión

Otra opción de diseño importante es qué teoría de la decisión debería utilizar la IA. Esto puede afectar a la manera en que la IA se comporte en ciertas situaciones estratégicamente decisivas. Podría determinar, por ejemplo, si la IA estaría abierta a negociar con, o ser extorsionada por otras civilizaciones superinteligentes cuya existencia hipotetice. Los detalles de la teoría de la decisión también podrían importar en predicamentos que implicaran probabilidades finitas de recompensas infinitas ("apuestas pascalianas") o probabilidades extremadamente pequeñas de recompensas finitas extremadamente grandes ("atracos pascalianos") o en contextos donde la IA se enfrentara a una incertidumbre normativa fundamental o donde hubiera múltiples instancias del mismo programa agente.[36]

Las opciones sobre la mesa incluyen la teoría causal de la decisión (en su variedad de sabores) y la teoría de la decisión evidencial, junto con otros candidatos más recientes, como la "teoría intemporal de la decisión" y la "teoría de la decisión no-actualizada", que todavía están en desarrollo.[37] Puede resultar difícil identificar y articular la teoría de la decisión correcta y tener una confianza justificada en que hayamos acertado. Aunque las perspectivas de poder especificar directamente la teoría de la decisión de una IA son quizás más optimistas que las de poder especificar directamente sus valores finales, aún nos enfrentamos a un sustancial riesgo de errar. Muchas de las complicaciones que podrían destruir las teorías de la decisión actualmente más populares fueron descubiertas recientemente, lo que sugiere que podrían existir más problemas que aún no hayamos

avistado. El resultado de dar a la IA una teoría de la decisión errónea podría ser desastroso, posiblemente de un calibre que ascendería a catástrofe existencial.

En vista de estas dificultades, podríamos considerar realizar un enfoque indirecto para especificar la teoría de la decisión que la IA debería utilizar. Cómo hacer esto exactamente aún no está claro. Podríamos querer que la IA usara "la teoría de la decisión D que hubiéramos querido que utilizara si hubiéramos pensado mucho sobre el asunto" Sin embargo, la IA tendría que ser capaz de tomar decisiones antes de aprender qué es D. Necesitaría, por lo tanto, algún tipo efectivo de teoría provisional de la decisión D' que guiara su búsqueda de D. Uno podría tratar de definir D' como una especie de superposición de las hipótesis actuales de la IA sobre D (sopesado por sus probabilidades), aunque hay problemas técnicos no resueltos para que esto se convirtiera en un camino totalmente genérico.[38] También es motivo de preocupación que la IA pudiera hacer malas decisiones irreversibles (como reescribirse a sí misma para ejecutarse a partir de ese momento en alguna teoría de la decisión defectuosa) durante la fase de aprendizaje, antes de que la IA hubiera tenido la oportunidad de determinar qué teoría de decisión particular es la correcta. Para reducir el riesgo de descarrilamiento durante este período de vulnerabilidad podríamos en su lugar tratar de dotar a la IA seminal con alguna forma de *racionalidad limitada*: una teoría de la decisión deliberadamente simplificada pero con un poco de suerte fiable que ignorara incondicionalmente consideraciones esotéricas, incluso las que pensáramos que podrían ser legítimas en última instancia, y que estuviera diseñada para reemplazarse a sí misma con una teoría de la decisión más sofisticada (e indirectamente especificada) una vez que ciertas condiciones se cumplie- ran.[39] Es una pregunta abierta a investigación si esto pudiera llegar a funcionar, y cómo.

Epistemología

Un proyecto también tendría que tomar una decisión fundamental de diseño en la selección de la epistemología de la IA, especificando los principios y criterios desde los cuáles las hipótesis empíricas deben ser evaluadas. Dentro de un marco bayesiano, podemos pensar en la epistemología como una función de probabilidad a priori —la asignación implícita de probabilidades de la IA a los mundos posibles antes de tomar ninguna evidencia perceptual en cuenta. En otros marcos, la epistemología podría tomar una forma diferente; pero en cualquier caso es necesaria alguna regla de aprendizaje inductivo si la IA fuera a generalizar a partir de observaciones pasadas y fuera a hacer predicciones sobre el futuro.[40] Sin embargo, al igual que con el

contenido de los objetivos y con la teoría de la decisión, existe el riesgo de que nuestra especificación epistemológica estuviera equivocada.

Uno podría pensar que hay un límite a cuánto daño podría surgir de una epistemología incorrectamente especificada. Si la epistemología fuera *demasiado* disfuncional, entonces la IA no podría ser muy inteligente y no podría plantear el tipo de riesgos que se analizan en este libro. Pero la preocupación es que pudiéramos especificar una epistemología que fuera suficientemente sólida como para hacer que la IA fuera instrumentalmente efectiva en la mayoría de situaciones, pero que tuviera algún defecto que desviara a la IA en algún asunto de crucial importancia. Tal IA podría ser similar a una persona perspicaz cuya cosmovisión se basara en un falso dogma, al que se aferrara con absoluta convicción, que en consecuencia "se enfrentara a molinos de viento" y lo diera todo en la búsqueda de objetivos fantasiosos o perjudiciales.

Ciertos tipos de diferencias sutiles en los principios de una IA podrían llegar a marcar una drástica diferencia en su comportamiento. Por ejemplo, una IA podría tener unos principios que asignaran una probabilidad cero a que el universo fuera infinito. No importa cuánta evidencia astronómica abogue por la tesis contraria, una IA de este tipo rechazaría obstinadamente cualquier teoría cosmológica que implicara un universo infinito; y podría tomar decisiones insensatas como resultado.[41] O una IA podría tener unos principios que asignaran una probabilidad cero a que el universo no fuera computable según Turing (esto es, de hecho, una característica común de muchos de los principios discutidos en literatura, incluyendo la complejidad de Kolmogorov previamente mencionada en el capítulo 1), de nuevo con consecuencias poco conocidas si el presupuesto enquistado —conocido como la "tesis de la Church-Turing"— resultara ser falsa. Una IA también podría terminar con unos principios que implicaran fuertes compromisos metafísicos de uno u otro tipo, por ejemplo, descartar a priori la posibilidad de que cualquier forma fuerte de dualismo mente-cuerpo pudiera ser verdadera, o la posibilidad de que haya hechos morales irreductibles. Si alguna de esas asunciones estuviera equivocada, la IA podría tratar de alcanzar sus objetivos finales de maneras que podríamos considerar como suplantaciones perversas. Sin embargo, no hay ninguna razón obvia por la que una IA tal, a pesar de estar fundamentalmente equivocada sobre una cuestión importante, no pudiera ser lo suficientemente instrumentalmente efectiva como para lograr una ventaja estratégica decisiva. (La antrópica, el estudio de cómo hacer inferencias a partir de la información indicativa en presencia de efectos de selección observacional, es otra área en la que la elección de los axiomas epistémicos podría resultar decisiva[42]).

Podríamos dudar razonablemente de nuestra capacidad para resolver todas las cuestiones fundamentales de epistemología para cuando construyamos la primera IA seminal. Podemos, por lo tanto, considerar la adopción de un enfoque indirecto en la tarea de especificar la epistemología de la IA. Esto conllevaría muchos de los mismos problemas que suscitaba el enfoque indirecto propuesto para especificar su teoría de la decisión. En el caso de la epistemología, sin embargo, puede haber una mayor esperanza de convergencia benigna, ya que cualquier epistemología de entre una amplia gama podría proporcionar un fundamento adecuado para una IA segura y eficaz, dando en última instancia, resultados doxásticos similares. La razón de esto es que unas pruebas suficientemente abundantes y un análisis empírico tenderían a enjuagar las diferencias moderadas en principios esperados.[43]

Un buen objetivo sería dotar a la IA con los mismos principios epistemológicos fundamentales que rigen nuestro propio pensamiento. Cualquier IA divergente de este ideal sería una IA que juzgaríamos que estaría razonando incorrectamente si aplicáramos constantemente nuestros propios estándares. Por supuesto, esto se aplica solamente a nuestros principios epistemológicos *fundamentales*. Los principios fundamentales deberían ser continuamente creados y revisados por la propia IA seminal mientras desarrolla su comprensión del mundo. El sentido de la superinteligencia no es mimar las preconcepciones humanas, sino hacer picadillo nuestra ignorancia y necedad.

Ratificación

El último punto en nuestra lista de opciones de diseño es la *ratificación*. ¿Deberían los planes de la IA ser objeto de revisión humana antes de ser puestos en práctica? Para un oráculo, esta pregunta se responde de manera implícita en su constitución. El oráculo da salida a información; los colaboradores humanos eligen cuándo y cómo actuar sobre ella. Para genios, soberanos, e IAs-herramientas, sin embargo, la cuestión de si se debe utilizar algún tipo de ratificación sigue abierta.

Para ilustrar cómo podría funcionar la ratificación, consideremos una IA destinada a funcionar como un soberano que implementara la VCE de la humanidad. En lugar de poner en marcha esta IA directamente, imaginemos que primero construimos una IA de tipo oráculo con el único propósito de responder a las preguntas acerca de lo que el soberano IA fuera a hacer. Como los capítulos anteriores revelaron, existen riesgos en la creación de un oráculo superinteligente (tales como los riesgos de crimen mental o de profusión infraestructural). Sin embargo, para efectos de este ejemplo

supongamos que el oráculo IA ha sido implementado con éxito de manera que evita estos escollos.

Tendríamos así un oráculo de IA que nos ofrecería sus mejores conjeturas acerca de las consecuencias de la ejecución de algún código destinado a aplicar la VCE de la humanidad. El oráculo podría no ser capaz de predecir en detalle lo que fuera a pasar, pero sus predicciones probablemente fueran mejores que las nuestras. (Si fuera imposible, incluso para una superinteligencia, predecir *nada* sobre qué haría el código, tendríamos que estar locos para ejecutarlo). Así que el oráculo reflexionaría durante un tiempo y luego presentaría su previsión. Para que la respuesta fuera inteligible, el oráculo podría ofrecer al operador una gama de herramientas con las que explorar diversas características de los resultados previstos. El oráculo podría mostrar imágenes de lo que el futuro parece y proporcionar estadísticas sobre el número de seres sintientes que existirán en diferentes momentos, junto con la media, el máximo, y los niveles más bajos de bienestar. Podría ofrecer biografías íntimas de varios individuos seleccionados al azar (quizá personas imaginarias seleccionadas por ser probablemente representativas). Podría destacar aspectos del futuro sobre los que el operador podría no haber pensado en preguntar, pero que serían considerados como pertinentes una vez fueran señalados.

Ser capaz de obtener una vista previa de los resultados de esta manera tendría obvias ventajas. La vista previa podría revelar las consecuencias de un error de diseño en las especificaciones o en el código fuente de un proyectado soberano. Si la bola de cristal mostrara un futuro en ruinas, podríamos desechar el código del soberano IA planificado y probar otra cosa. Podría argumentarse con fuerza que deberíamos familiarizarnos con las ramificaciones concretas de una opción antes de comprometernos con ella, especialmente cuando todo el futuro de la raza humana esté en juego.

Lo que quizás es menos obvio es que la ratificación tiene también desventajas potencialmente significativas. La calidad conciliadora de la VCE podría ser socavada si facciones opuestas, en lugar de someterse al arbitraje de una sabiduría superior, esperando y confiando en ser vengadas, pudieran ver de antemano cuál sería el veredicto. Un defensor del enfoque basado en la moralidad podría preocuparse de que la resolución del patrocinador se derrumbaría si todos los sacrificios requeridos por la moral óptima llegaran a ser revelados. Y todos podríamos tener razón para preferir un futuro que tuviera sorpresas, disonancias, un toque salvaje, oportunidades para la auto-superación —un futuro cuyos contornos no estuvieran demasiado ajustados a ideas preconcebidas, sino que ofrecieran algún movimiento dramático y crecimiento inesperado. Podríamos ser

menos propensos a tener una visión tan expansiva si pu-

diéramos elegir cada detalle del futuro, enviando de vuelta cualquier proyecto que no se ajustara totalmente a nuestra fantasía en ese momento.

Por tanto, la cuestión de la ratificación del patrocinador es menos clara de lo que inicialmente podría parecer. No obstante, parece prudente, a fin de cuentas, aprovechar la oportunidad de previsualización, si esa funcionalidad estuviera disponible. Pero en lugar de dejar que el revisor afinara todos los aspectos de los resultados, podríamos darle un veto sencillo que sólo pudiera ser ejercido un par de veces antes de que todo el proyecto fuera abortado.[44]

Acercándonos lo suficiente

El objetivo principal de la ratificación sería reducir la probabilidad de un error catastrófico. En general, parece prudente apuntar a minimizar el riesgo de error catastrófico en lugar de maximizar la probabilidad de que cada detalle esté totalmente optimizado. Hay dos razones para ello. En primer lugar, los recursos cósmicos de la humanidad son astronómicamente grandes, —hay suficiente para todos, incluso si nuestro proceso trajera consigo algunos residuos o aceptara algunas restricciones innecesarias. En segundo lugar, existe la esperanza de que sólo con conseguir las condiciones iniciales aproximadamente correctas para la explosión de inteligencia, la superinteligencia resultante podría llevar todo adelante y acertar con nuestros objetivos finales. Lo importante es que aterrice en el lugar adecuado.

Con respecto a la epistemología, es plausible que una amplia gama de principios en última instancia converjan en resultados muy similares (cuando fueran calculados por una superinteligencia y condicionados por una cantidad realista de datos). Por lo tanto, no hay necesidad de preocuparse por obtener una epistemología *exactamente* correcta. Debemos simplemente no dar a la IA unos principios que fueran tan extremos como para hacer que la IA fuera incapaz de aprender verdades vitales, incluso con el beneficio de mucha experiencia y análisis.[45]

Con respecto a la teoría de la decisión, el riesgo de error irrecuperable parece más grande. Todavía podríamos esperar poder especificar directamente una teoría de la decisión que fuera suficientemente buena. Una IA superinteligente podría cambiarse a una nueva teoría de la decisión en cualquier momento; sin embargo, si se comienza con una teoría de la decisión suficientemente mala podría no llegar a ver razones para cambiar. Incluso si un agente llegara a ver

los beneficios de tener una teoría de la decisión diferente, su plasmación podría llegar demasiado tarde. Por ejemplo, un agente diseñado para rechazar el chantaje podría gozar del beneficio de disuadir a posibles ex- torsionistas. Por esta razón, los agentes chantajeables harían bien en adoptar de forma proactiva una teoría de la decisión inexplotable. Sin embargo, una vez que un agente chantajeable recibiera la amenaza y la considerara creíble, el daño estaría hecho.

Dada una epistemología y una teoría de la decisión adecuadas, podríamos tratar de diseñar el sistema que implementara la VCE o algún otro tipo de contenido objetivo especificado indirectamente. Una vez más habría esperanza de convergencia: que las diferentes formas de implementar una dinámica similar a la VCE conducirían al mismo resultado utópico. A falta de dicha convergencia, todavía podría esperarse que muchos de los diferentes resultados posibles fueran suficientemente buenos como para contar como éxitos existenciales.

No es necesario que creemos un diseño altamente optimizado. Más bien, nuestra atención debe centrarse en la creación de un diseño altamente confiable, uno en el que podamos confiar en que mantendrá suficiente cordura como para reconocer sus propios defectos. Una superinteligencia imperfecta, cuyos fundamentos sean sólidos, se repararía gradualmente a sí misma; y, una vez lo hubiera hecho, ejercería tanto poder de optimización beneficioso sobre el mundo como si hubiera sido perfecta desde el principio.

CAPÍTULO 14

El panorama estratégico

A

hora es el momento de considerar el desafío de la superinteligencia en un contexto más amplio. Nos gustaría orientarnos en el panorama estratégico lo suficiente como para saber por lo menos qué dirección general deberíamos estar tomando. Esto, por lo que parece, no es nada fácil. Aquí, en el penúltimo capítulo, introduciremos algunos conceptos analíticos generales que nos ayudarán a reflexionar sobre cuestiones de política científica y tecnológica a largo plazo. A continuación, los aplicaremos al tema de la inteligencia artificial.

Puede ser esclarecedor hacer una distinción aproximada entre dos posturas normativas diferentes desde las que se puede evaluar una propuesta. *La perspectiva de la persona afectada* se pregunta si el cambio propuesto iría en "nuestro interés" —es decir, si iría (en general, y en cuanto a sus expectativas) en el interés de esas criaturas moralmente susceptibles de consideración que, o bien ya existen, o llegarán a existir con independencia de que el cambio propuesto se produzca o no. *La perspectiva impersonal,* por el contrario, no da ninguna consideración especial a las personas existentes en la actualidad, ni a los que llegarán a existir independientemente de que se produzca el cambio propuesto. En su lugar, toma en cuenta a todos por igual, independientemente de su ubicación temporal. La perspectiva impersonal da un gran valor a traer gente nueva a la

existencia, siempre que vayan a tener una vida que valiera la pena vivir: cuantas más vidas felices sean creadas, mejor.

Esta distinción, aunque apenas aluda a las complejidades morales asociadas con una revolución de inteligencia artificial, puede ser útil en un primer análisis. Aquí examinaremos primero las cuestiones desde la perspectiva impersonal. Más adelante veremos qué cambia si se le da peso en nuestras deliberaciones a la perspectiva de personas afectadas.

Estrategia científica y tecnológica

Antes de acercarnos a cuestiones específicas de la superinteligencia artificial, debemos introducir algunos conceptos estratégicos y consideraciones que se refieren al desarrollo científico y tecnológico en general.

Desarrollo tecnológico diferencial

Supongamos que un político propone recortar los fondos para un campo de investigación determinado, debido a la preocupación por los riesgos o consecuencias a largo plazo de una hipotética tecnología que eventualmente podría surgir de dicha investigación. Este político se encontrará con una gran oposición por parte de la comunidad investigadora.

Los científicos y sus defensores públicos a menudo dicen que es inútil tratar de controlar la evolución tecnológica bloqueando la investigación. Si alguna tecnología es factible (afirma su argumento), se desarrollará independientemente de los escrúpulos que cualquier autoridad normativa tenga sobre sus posibles riesgos futuros. De hecho, cuanto más potentes sean las capacidades que una línea de desarrollo prometa producir, más seguros podemos estar de que alguien, en algún lugar, estará motivado para conseguirlo. Los recortes de fondos no detendrán el progreso ni prevendrán contra sus peligros concomitantes.

Curiosamente, esta objeción de inutilidad casi nunca se plantea cuando un político se propone *aumentar* la financiación para algún área de investigación, a pesar de que el argumento parece funcionar en ambos sentidos. Rara vez se oye voces indignadas protestar: “Por favor, no aumenten nuestros fondos. En su lugar, hagan algunos recortes. Los investigadores de otros países seguramente tomarán el relevo; el mismo trabajo será hecho de todos modos. ¡No malgasten el dinero público en investigación científica nacional!”

¿Cómo se explica esta aparente ambigüedad en su planteamiento? Una explicación plausible, por supuesto, es que los miembros de la comunidad de investigación tenemos un sesgo egoísta que nos lleva a

creer que la investigación es siempre buena y estamos tentados de aceptar casi cualquier argumento que apoye nuestra demanda de más fondos. Sin embargo, también es posible que la doble moral se pueda justificar en términos de interés propio nacional. Supongamos que el desarrollo de una tecnología tiene dos efectos: darle un pequeño beneficio B a sus inventores y al país que les patrocina, mientras que la imposición de un daño agregado mayor H —que podría ser un riesgo externo— a todo el mundo. Incluso alguien que fuera bastante altruista podría optar en este caso por desarrollar la tecnología dañina. Puede que razonen que el daño H tendrá lugar sin importar lo que hagan, ya que si se abstienen alguien más desarrollaría la tecnología de todos modos; y dado que el bienestar total no puede ser afectado, mejor conseguir el beneficio B para ellos y su nación. ("Desgraciadamente, pronto habrá un dispositivo que destruirá el mundo. Afortunadamente, ¡tenemos el honor de construirlo!").

Cualquier explicación de la apelación a la objeción de inutilidad no logra demostrar que no haya, en general, razones impersonales para tratar de dirigir el desarrollo tecnológico. No lo logra incluso si concedemos la idea motivadora de que con los continuos esfuerzos de desarrollo científico y tecnológico, todas las tecnologías relevantes se desarrollarían —es decir, incluso si concedemos lo siguiente:

Conjetura de la compleción tecnológica

Si los esfuerzos en desarrollo científico y tecnológico no cesan de manera efectiva, entonces

todas las capacidades básicas importantes que podrían ser obtenidas a través de alguna posible

tecnología, serán obtenidas.[1]

Hay al menos dos razones por las que la conjetura de compleción tecnológica no implica la objeción de futilidad. En primer lugar, el antecedente podría no sostenerse, porque no es, de hecho, algo dado el que los esfuerzos de desarrollo científico y tecnológico no vayan a cesar efectivamente (antes de la consecución de la madurez tecnológica). Esta reserva es especialmente pertinente en un contexto que implique un riesgo existencial. En segundo lugar, incluso si pudiéramos estar seguros de que se obtendrían todas las capacidades básicas importantes que pudieran ser obtenidas a través de alguna posible tecnología, podría todavía tener sentido tratar de influir en la dirección de la investigación tecnológica. Lo que importa no es sólo *si* una tecnología se desarrolla, sino también *cuándo* se desarrolla, por *quién* y en *qué contexto*. Estas circunstancias del nacimiento de una nueva tecnología, que dan forma a su impacto, pueden verse afectadas abriendo o cerrando grifos de financiación (y mediante otros instrumentos políticos).

Estas reflexiones sugieren un principio que nos haga atender a la velocidad relativa con la que se desarrollan las diferentes tecnologías:[2]

> **El principio de desarrollo tecnológico diferencial**
> Retrasar el desarrollo de tecnologías peligrosas y perjudiciales, especialmente las que aumenten el nivel de riesgo existencial; y acelerar el desarrollo de tecnologías beneficiosas, especialmente aquellas que reduzcan los riesgos existenciales planteados por la naturaleza o por otras tecnologías.

Una política podría entonces evaluarse sobre la base de cuánta ventaja diferencial da a las formas deseables de desarrollo tecnológico frente a las formas indeseables.[3]

Preferencia en el orden de llegada

Algunas tecnologías tienen un efecto ambivalente sobre los riesgos existenciales, aumentando algunos riesgos existenciales, mientras que disminuyen otros. La superin- teligencia es una de estas tecnologías.

Hemos visto en capítulos anteriores que el surgimiento de superinteligencia artificial crearía un sustancial riesgo existencial. Pero también reduciría muchos otros riesgos existenciales. Los riesgos provenientes de la naturaleza —tales como los impactos de asteroides, los supervolcanes y las pandemias naturales— serían virtualmente eliminados, ya que la superinteligencia podría implementar contramedidas contra la mayoría de riesgos de este tipo, o al menos disminuir el nivel a la categoría de no-exis- tencial (por ejemplo, a través de la colonización del espacio).

Estos riesgos existenciales provenientes de la naturaleza son

relativamente pequeños en grandes períodos temporales. Pero la superinteligencia también eliminaría o reduciría muchos riesgos antropogénicos. En particular, reduciría los riesgos de destrucción accidental, incluyendo el riesgo de que ocurrieran accidentes relacionados con las nuevas tecnologías. Al ser, en general, más capaces que los seres humanos, una superinteligencia sería menos propensa a cometer errores, y tendría más probabilidades de reconocer cuando sería necesario tomar precauciones y ponerlas en práctica de manera competente. Una superinteligencia bien construida podría llegar a tomar riesgos, pero sólo cuando hacerlo fuera aconsejable. Además, al menos en escenarios donde la superinteligencia formara una Unidad, muchos riesgos existenciales an

tropogénicos no accidentales derivados de problemas de coordinación global serían eliminados. Estos riesgos incluyen el riesgo de guerras, las carreras tecnológicas, las formas indeseables de competencia y evolución, y las tragedias de los comunes.

Puesto que habría un peligro sustancial asociado a que los seres humanos desarrollaran la biología sintética, la nanotecnología molecular, la ingeniería climática, los instrumentos de mejora biomédica y manipulación neuropsicológica, herramientas de control social que podrían facilitar el totalitarismo o la tiranía, y otras tecnologías hasta ahora inimaginables, eliminar estos riesgos debería ser muy beneficioso. Por tanto, podría argumentarse que cuanto antes llegara la superinteligencia, mejor. No obstante, si los riesgos provenientes de la naturaleza y de otros peligros no relacionados con la tecnología del futuro son pequeños, entonces este argumento podría ser refinado: lo importante sería alcanzar la superinteligencia *antes* que otras tecnologías peligrosas, como la nanotecnología avanzada. Que ocurra tarde o temprano puede no ser tan importante (desde una perspectiva impersonal) como que el orden de llegada sea el correcto.

La base para preferir que la superinteligencia llegue antes que otras tecnologías potencialmente peligrosas, como la nanotecnología, es que la superinteligencia reduciría los riesgos existenciales de la nanotecnología, pero no viceversa.[4] Por lo tanto, si creamos primero la superinteligencia, nos enfrentaremos sólo los riesgos existenciales asociados a la superinteligencia; mientras que si creamos primero la nanotecnología, nos enfrentaremos a los riesgos de la nanotecnología y luego, adicionalmente, a los riesgos de la superinteligencia.[5] Incluso si los riesgos existenciales de la superinteligencia son muy grandes, e incluso si la superinteligencia es la más peligrosa de todas las tecnologías, podría defenderse el acelerar su llegada.

Estos argumentos de "cuanto-antes-mejor", sin embargo, presuponen que el grado de riesgo que conlleva crear una superinteligencia es el mismo con independencia del momento de su creación. Si, en cambio, su grado de riesgo disminuyera con el tiempo, tal vez sería mejor retrasar la revolución de la inteligencia artificial. Mientras que una llegada tardía daría más tiempo para que otras catástrofes existenciales sucedieran, incluso en ese caso podría ser preferible retrasar el desarrollo de la super- inteligencia. Esto sería especialmente plausible si los riesgos existenciales asociados con la superinteligencia fueran mucho mayores que los asociados a otras tecnologías disruptivas.

Hay varias razones muy fuertes para creer que el grado de peligro de una explosión de inteligencia se reducirá significativamente a lo largo de un plazo de varias décadas. Una razón es que una fecha posterior deja más tiempo para el desarrollo de soluciones para el

problema del control. El problema del control ha sido reconocido recientemente, y la mayoría de las mejores ideas sobre cómo afrontarlo fueron descubiertas en la última década aproximadamente (y en varios casos mientras este libro se estaba escribiendo). Es posible que el estado de la cuestión avance mucho en las próximas décadas; y si el problema resulta ser muy difícil, una importante tasa de progreso podría mantenerse durante un siglo o más. Cuanto más tiempo se tarde en llegar a la superinteligencia, más progreso habremos alcanzado cuando llegue. Esta es una consideración importante a favor de retrasar la fecha de llegada —y una fuerte consideración en contra de adelantar la fecha de llegada.

Otra razón por la que podría ser más seguro retrasar la superinteligencia es que esto podría dar más tiempo a que las diversas tendencias beneficiosas de la civilización humana se desarrollaran. Cuánto peso se otorgue a esta consideración dependerá de lo optimistas que seamos respecto de estas tendencias.

Un optimista sin duda podría apuntar a una serie de alentadores indicadores y posibilidades esperanzadoras. La gente puede aprender a llevarse mejor, lo que llevaría reducciones de la violencia, la guerra y la crueldad; y la coordinación global y el alcance de la integración política podrían aumentar, por lo que sería más fácil escapar de carreras tecnológicas indeseables (más sobre esto a continuación) y llegar a un acuerdo por el cual las ganancias esperadas de una explosión de inteligencia fueran ampliamente compartidas. Parece que hay tendencias históricas de larga duración que apuntan en estas direcciones.[6]

Además, un optimista podría esperar que el "nivel de cordura" de la humanidad se elevara en el transcurso de este siglo —que los prejuicios (en su conjunto) se redujeran, que los conocimientos se acumularan, y que la gente se acostumbrara más a pensar en las abstractas probabilidades futuras y en los riesgos globales. Con suerte, podríamos ver un aumento general de los estándares epistémicos tanto en la cognición individual como en la colectiva. Una vez más, hay tendencias que empujan en esa dirección. El progreso científico significa que se sabrán más cosas. El crecimiento económico podría dar alimentación adecuada a una mayor parte de la población mundial (sobre todo durante los primeros años de la vida que tan importantes son para el desarrollo cerebral) y acceso a una educación de calidad. Los avances en la tecnología de la información harán que sea más fácil encontrar, integrar, evaluar y comunicar datos e ideas. Por otra parte, a finales de siglo, la humanidad tendrá en su haber cien años más de errores, de los que podrá aprender.

Muchos desarrollos potenciales son ambivalentes en el sentido antes mencionado —aumentan algunos riesgos existenciales y

disminuyen otros. Por ejemplo, los avances en vigilancia, la minería de datos, la detección de mentiras, la biometría, y los medios psicológicos o neuroquímicos para manipular las creencias y los deseos podrían reducir algunos riesgos existenciales coordinándolos más fácilmente a nivel internacional o reprimiendo a terroristas y rebeldes. Estos mismos avances, sin embargo, también podrían aumentar algunos riesgos existenciales amplificando dinámicas sociales indeseables o permitiendo la formación de regímenes totalitarios permanentemente estables.

Una frontera importante es la mejora de la cognición biológica a través de la selección genética. Cuando hablamos de esto en los capítulos 2 y 3, llegamos a la conclusión de que sería más probable que las formas más radicales de superinteligencia surgieran como inteligencia artificial. Esa afirmación es coherente con la idea de que la mejora cognitiva jugará un papel importante en el período previo a la creación de la superinteligencia artificial. La mejora cognitiva puede parecer obviamente reductora de riesgos: cuanto más inteligentes sean las personas trabajando en el problema de control, más probabilidades hay de encontrar una solución. Sin embargo, la mejora cognitiva también podría acelerar el desarrollo de la inteligencia artificial, lo que reduciría el tiempo disponible para trabajar en el problema. La mejora cognitiva también tendría otras muchas consecuencias relevantes. Estas cuestiones merecen una mirada más cercana. (La mayoría de las siguientes observaciones sobre la "mejora cognitiva" se aplican igualmente a los medios no biológicos para incrementar nuestra efectividad espistémica individual o colectiva).

El ritmo de cambio y la mejora cognitiva

Un aumento en la media o en el rango superior de la capacidad intelectual humana probablemente acelere el progreso tecnológico en todos los ámbitos, incluyendo el progreso hacia diversas formas de la inteligencia artificial, el progreso en el problema de control, y el progreso en una amplia franja de otros objetivos técnicos y económicos. ¿Cuáles serían las consecuencias netas de tal aceleración?

Consideremos el caso límite de un "acelerador universal", una intervención imaginaria que acelerara literalmente *todo*. La acción de un acelerador universal tal, simplemente conllevaría un cambio arbitrario en la periodización del tiempo, sin producir ningún cambio cualitativo en los resultados observados.[7]

Si hemos de dar sentido a la idea de que la mejora de la cognición podría, por lo general, acelerar las cosas, necesitamos claramente algún otro concepto más allá distinto al de aceleración universal. Un enfoque más prometedor es centrarse en cómo la mejora cognitiva

podría aumentar el ritmo de cambio en un tipo de proceso en relación al ritmo de cambio en algún otro tipo de proceso. Tal aceleración diferencial podría afectar las dinámicas de un sistema. Por lo tanto, consideremos el siguiente concepto:

Acelerador de desarrollo macroestructural - una palanca que acelera el ritmo en el que las características macro-estructurales de la condición humana se desarrollan, sin cambiar el ritmo en que se desarrollan los asuntos humanos a nivel micro.

Imagínese tirando de esta palanca en la dirección de desaceleración. Una pastilla de freno aprieta la gran rueda de la historia del mundo; saltan chispas y crujidos metálicos. Cuando la rueda se hubiera adaptado a un ritmo más pausado, el resultado sería un mundo en el que la innovación tecnológica se produciría más lentamente y en el que los cambios fundamentales o de importancia mundial en la estructura y cultura política sucederían con menos frecuencia y de manera menos abrupta. Pasarían un mayor número de generaciones antes de que una época diera paso a otra. Durante el curso de una vida útil, una persona vería pocos cambios en la estructura básica de la condición humana.

Durante la mayor parte de la existencia de nuestra especie, el desarrollo macro-es- tructural fue más lento de lo que es ahora. Hace cincuenta mil años, todo un milenio podría haber transcurrido sin una sola invención tecnológica significativa, sin ningún aumento notable en el conocimiento y el entendimiento humano, y sin ningún cambio político a nivel mundial significativo. En un nivel micro, sin embargo, el caleidoscopio de los asuntos humanos se agitaría a un ritmo razonable, con nacimientos, muertes y otros eventos personal y localmente importantes. El día de la persona promedio podría haber sido más agotador en el Pleistoceno de lo que es hoy.

Si nos encontráramos con una palanca mágica que permitiera cambiar el ritmo de desarrollo macroestructural, ¿qué deberíamos hacer? ¿deberíamos acelerar, desacelerar o dejar las cosas como estuvieran?

Asumiendo un punto de vista impersonal, esta pregunta nos obliga a considerar las posibilidades de riesgo existencial. Distingamos entre dos tipos de riesgo: los “riesgos de estado” y los “riesgos de transición”. Un riesgo de estado es uno que está asociado a estar en un cierto estado, y la cantidad total de riesgo del estado al que un sistema está expuesto es una función directa de cuánto tiempo se mantenga el sistema en ese estado. Los riesgos de la naturaleza son típicamente riesgos de estado: cuanto más tiempo permanecemos expuestos, mayor es la probabilidad de que vamos a ser alcanzados por un asteroide, una supererupción volcánica, un estallido de rayos gamma, una pandemia que surja de forma natural, o alguna otra variedad de guadaña cósmica. Algunos riesgos antropogénicos son también riesgos

de estado. A nivel individual, cuanto más tiempo asome un soldado la cabeza por encima del parapeto, mayor es la probabilidad acumulativa de que sea disparado por un francotirador enemigo. También hay riesgos de estado antropogénicos a nivel existencial: cuanto más tiempo vivamos en un sistema internacional anárquico, mayor es la probabilidad acumulativa de un Armagedón termonuclear o de una gran guerra en la que se empleen otros tipos de armas de destrucción masiva, arrasando la civilización.

Un riesgo de transición, por el contrario, es un riesgo discreto asociado con algún tipo de transición necesaria o deseable. Una vez que se completa la transición, el riesgo desaparece. La cantidad de riesgo de transición asociado con una transición, por lo general, no es una simple función de cuánto tiempo tarda la transición. No se reduce a la mitad el riesgo de atravesar un campo minado haciéndolo el doble de rápido. Si se da un despegue rápido, la creación de superinteligencia podría ser un riesgo de transición: habría un cierto riesgo asociado con el despegue, cuya magnitud dependerá de los preparativos que se hayan hecho; pero la cantidad de riesgo podría no depender tanto de si el despegue tarda veinte milisegundos o veinte horas.

Entonces podemos decir lo siguiente con respecto a un hipotético acelerador de desarrollo macro-estructural:

- En la medida en que estamos preocupados con riesgos existenciales de estado, debemos favorecer la aceleración —siempre que pensemos tener una perspectiva realista de llegar a una era de post-transición en la que futuros riesgos existenciales se reducirían considerablemente.
- Si se supiera por adelantado que habría algo destinado a causar una catástrofe existencial, entonces deberíamos reducir el ritmo de desarrollo macro-estructural (o incluso dar marcha atrás) con el fin de dar la oportunidad de existir a más generaciones antes de que se bajara el telón. Pero sería demasiado pesimista estar seguro de que, de hecho, la humanidad está condenada.
- En la actualidad, el nivel de riesgo de estado existencial parece ser relativamente bajo. Si imaginamos las macro-condiciones tecnológicas de la humanidad congeladas en su estado actual, parece muy poco probable que se produzca una catástrofe existencial en el lapso de, digamos, una década. Así que un retraso de una década —siempre que ocurriera en nuestra etapa actual de desarrollo o en algún otro momento en el que el riesgo de estado fuera bajo— conllevaría sólo un riesgo de estado existencial muy pequeño, mientras que aplazar una década ciertos desarrollos tecnológicos podría tener un efecto beneficioso significativo sobre los riesgos existenciales de transición posteriores al permitir por ejemplo, más tiempo para la preparación.

Conclusión: la principal forma en que el ritmo de desarrollo macro-estructural podría ser importante sería si afectara a lo preparada que la humanidad estuviera cuando llegara el momento de enfrentarse a los riesgos de transición claves.[8]

Así que la pregunta que debemos hacernos es cómo afectaría la mejora cogniti- va (y la aceleración asociada al desarrollo macro-estructural) al nivel esperado de preparación en el momento crítico. ¿Deberíamos preferir un período más corto de preparación con una inteligencia superior? Con una inteligencia superior, el tiempo de preparación podría utilizarse de manera más eficaz, y el paso crítico final sería dado por una humanidad más inteligente. ¿O deberíamos preferir operar con niveles de inteligencia cercanos a los actuales, si eso nos da más tiempo para prepararnos?

Qué opción sea mejor depende de la naturaleza del desafío para el que nos estemos preparando. Si el reto consistiera en resolver un problema en el que aprender de la experiencia fuera la clave, entonces la longitud cronológica del período de preparación podría ser el factor determinante, ya que se necesitaría tiempo para acumular la experiencia necesaria. ¿Cómo sería un desafío de este tipo? Un ejemplo hipotético sería una nueva tecnología de armamento que podríamos predecir sería desarrollada en algún momento futuro y que supondría que cualquier guerra posterior tendría, digamos, una probabilidad de uno entre diez de causar una catástrofe existencial. Si el desafío al que nos enfrentáramos fuera de esta naturaleza, entonces podríamos desear que el ritmo de desarrollo macro-estructural fuera lento, para que nuestra especie tuviera más tiempo de trabajar juntos antes de la etapa crítica cuando la nueva tecnología de armamento se inventara. Podríamos esperar que durante el período de gracia ganado a través de la desaceleración, nuestra especie pudiera aprender a evitar la guerra —que las relaciones internacionales de todo el mundo pudieran llegar a parecerse a las que existen entre los países de la Unión Europea, que, después de haber luchado entre sí ferozmente durante siglos, conviven ahora en relativa paz y armonía. La pacificación podría ocurrir como resultado de una edificación moderada desde diversos procesos civilizadores o a través de una terapia de choque a través de golpes sub-exis- tenciales (por ejemplo, pequeñas conflagraciones nucleares, y el retraimiento y resolución que podrían generar para crear finalmente las instituciones globales necesarias para abolir guerras interestatales). Si este tipo de aprendizaje o ajuste no se viera muy acelerado por el aumento de la inteligencia, entonces la mejora cognitiva sería indeseable, pues simplemente serviría para quemar rápidamente los fusibles.

Una explosión de inteligencia inminente, sin embargo, puede presentar un desafío de diferente tipo. El problema de control exige previsión, razonamiento y comprensión teórica. No está tan claro cómo ayudaría el aumento de experiencia histórica. La experiencia directa de una explosión de inteligencia no es posible (hasta demasiado tarde), y muchas características conspiran para hacer que

el problema de control sea único y carezca de precedentes históricos relevantes. Por estas razones, la cantidad de tiempo que transcurra antes de la explosión de inteligencia puede no importar mucho per se. Tal vez lo que importa, en cambio, es: (a) la cantidad de progreso intelectual sobre el problema de control logrado en el momento de la detonación; y (b) la cantidad de habilidad e inteligencia disponible en el momento de implementar las mejores soluciones disponibles (y de improvisar lo que falte).[9] Que este último factor debería estar favorecido por la mejora cognitiva es obvio. Cómo afectaría la mejora cognitiva al factor (a) es una cuestión un poco más sutil.

Supongamos, como se sugirió anteriormente, que la mejora cognitiva fuera un acelerador de desarrollo macro-estructural general. Esto aceleraría la llegada de la explosión de inteligencia, lo que reduciría la cantidad de tiempo disponible para prepararse y avanzar en el problema de control. Normalmente esto sería algo negativo. Sin embargo, si la única razón por la que habría menos tiempo disponible para el progreso intelectual es porque estaríamos acelerando el progreso intelectual, entonces no tendría por qué haber una reducción neta de la cantidad de progreso intelectual logrado antes de producirse la explosión de inteligencia.

En este punto, la mejora cognitiva podría parecer neutral con respecto al factor (a): el mismo progreso intelectual que de otra manera se hubiera hecho antes de la explosión de inteligencia —incluyendo el progreso en problema de control— estaría igualmente hecho, sólo que comprimido en un intervalo de tiempo más corto. En realidad, sin embargo, la mejora cognitiva podría llegar a tener una influencia positiva en (a).

Una razón por la cual la mejora cognitiva podría lograr más progreso sobre el problema de control para el momento en que la explosión de inteligencia ocurriera, es que el progreso en el problema de control puede estar especialmente supeditado a niveles extremos de rendimiento intelectual —incluso más que el tipo de trabajo necesario para crear inteligencia artificial. El papel del ensayo y error y de la acumulación de resultados experimentales parece bastante limitado en relación al problema de control, mientras que el aprendizaje experimental probablemente juegue un papel importante en el desarrollo de la inteligencia artificial o en la emulación de cerebro completo. La medida en que el tiempo pueda sustituir al ingenio puede, por tanto, variar entre las tareas, de manera que la mejora cognitiva podría promover un mayor progreso en el problema de control que en el problema de cómo crear inteligencia artificial.

Otra razón por la que la mejora cognitiva debería promover avances decisivos en el problema de control es que la propia necesidad de tal progreso es probable que sea más apreciada por sociedades e

individuos cognitivamente más capaces. Se requiere previsión y razonamiento para comprender por qué el problema de control es importante y para que sea un prioridad.[10] También puede ser necesaria una inusual sagacidad para encontrar maneras prometedoras de aproximarse a un problema tan poco familiar.

A partir de estas reflexiones podemos concluir tentativamente que la mejora cog- nitiva es deseable, al menos en la medida en que la atención se centre en los riesgos existenciales de una explosión de inteligencia. Líneas paralelas de pensamiento se aplican a otros riesgos existenciales que surgen a propósito de desafíos en los que son necesarias la previsión y el razonamiento abstracto fiable (a diferencia de, por ejemplo, la adaptación gradual a los cambios experimentados en el entorno o el proceso multigeneracional de maduración cultural y creación de instituciones).

Acoplamientos tecnológicos

Supongamos que pensáramos que la solución del problema de control para la inteligencia artificial fuera muy difícil, que la solución para las emulaciones de cerebro completo fuera mucho más fácil, y que, por lo tanto, sería preferible que se llegara a la inteligencia artificial a través de la ruta de la emulación de cerebro. Volveremos más tarde a la cuestión de si la emulación de cerebro completo sería más segura que la inteligencia artificial. Mas por ahora queremos dejar dicho que, incluso si aceptamos esta premisa, de ella no se sigue que debamos promover la tecnología de emulación de cerebro completo. Una de las razones, señalada anteriormente, es que una llegada tardía de la superinteligencia puede ser preferible, dando así más tiempo para avanzar en el problema de control y para que otras tendencias favorables puedan desarrollarse —y, por lo tanto, si uno estuviera seguro de que la emulación de cerebro completo precedería a la IA de todos modos, sería contraproducente acelerar aún más la llegada de la emulación de cerebro completo.

Pero incluso si se diera el caso de que lo mejor fuera que la emulación de cerebro completo llegara lo antes posible, *todavía* no se seguiría que debiéramos favorecer el progreso hacia la emulación de cerebro completo. Pues es posible que el progreso hacia la emulación de cerebro completo no nos lleve a la emulación de cerebro completo. Puede que en su lugar origine inteligencia artificial neuromórfica —formas de IA que imiten algunos aspectos de la organización cortical pero que no replican la funcionalidad neuronal con suficiente fidelidad como para constituir una emulación apropiada. Si —como hay razones para creer— tales IAs neuromórficas fueran peores que la clase de IAs se habrían construido en otro caso, y si mediante la

promoción de la emulación de cerebro completo hiciéramos que una IA neuromórfica llegara primero, entonces nuestra búsqueda del *mejor* resultado (la emulación de cerebro completo) conduciría al *peor* resultado (la IA neuromórfica); mientras que si hubiéramos perseguido *el segundo mejor* resultado (la IA sintética) podríamos realmente haber alcanzado el segundo mejor resultado (la IA sintética).

Acabamos de describir una (hipotética) instancia de lo que podríamos llamar un "acoplamiento tecnológico".[11] Esto se refiere a la situación en la que dos tecnologías tienen una relación temporal predecible, de tal manera que el desarrollo de una de las tecnologías tiene grandes posibilidades de llevar al desarrollo de la otra, ya sea como un precursor necesario o como una aplicación obvia e irresistible o como un paso subsiguiente. Los acoplamientos tecnológicos deben tenerse en cuenta cuando utilizamos el principio de desarrollo tecnológico diferencial: no es bueno acelerar el desarrollo de una tecnología deseable Y si la única manera de conseguir Y es mediante el desarrollo de una tecnología precursora extremadamente indeseable X, o si conseguir Y produciría inmediatamente una tecnología relacionada Z extremadamente indeseable. Antes de casarte con tu amor, toma en consideración a tu futura familia política.

En el caso de la emulación de cerebro completo, el grado de acoplamiento tecnológico es discutible. Hemos observado en el capítulo 2 que mientras que la emulación de cerebro completo requeriría avances masivos en varias tecnologías de apoyo, puede que no requiriera ninguna gran visión teórica nueva. En particular, no requiere que entendamos cómo funciona la cognición humana, sólo que sepamos cómo construir modelos computacionales de pequeñas partes del cerebro, tales como diferentes especies de neuronas. Sin embargo, en el curso del desarrollo de la capacidad de emulación del cerebro humano, una gran cantidad de datos neuroanatómicos serían almacenados, y los modelos funcionales de las redes corticales seguramente se mejorarían considerablemente. Tal progreso parece tener una buena oportunidad de posibilitar una IA neuromórfica antes de que tengamos una plena emulación de cerebro com- pleto.[12] Históricamente, hay un buen número de ejemplos de técnicas de IA tomadas de la neurociencia o de la biología. (Por ejemplo: la neurona de McCulloch-Pitts, los perceptrones y otras neuronas y redes neuronales artificiales, inspiradas en el trabajo neuroanatómico; el aprendizaje por refuerzo, inspirado en la psicología conductista; los algoritmos genéticos, inspirados en la teoría de la evolución; las arquitecturas de subsunción y las jerarquías perceptivas, inspirados por las teorías cognitivas científicas sobre la planificación motora y la percepción sensorial; los sistemas inmunológicos artificiales, inspirado

en la inmunología teórica; la mente-enjambre, inspirada en la ecología de las colonias de insectos y otros sistemas de auto-organización; y el control reactivo y basado en el comportamiento para la robótica, inspirados en el estudio de la locomoción animal). Tal vez lo más importante es que hay un montón de importantes preguntas relevantes para la IA que podrían ser respondidas mediante un mayor estudio del cerebro. (Por ejemplo: ¿Cómo guarda cerebro las representaciones estructuradas en la memoria de trabajo y en la memoria a largo plazo? ¿Cómo se resuelve el problema de unión? ¿Qué es el código neuronal? ¿Cómo se representan los conceptos? ¿Hay alguna unidad estándar de maquinaria de procesamiento cortical, como la columna cortical? Y, si es así, ¿cómo está cableada y cómo depende su funcionalidad del cableado? ¿Cómo pueden estas columnas enlazarse, y cómo pueden aprender?)

En breve tendremos más que decir sobre el peligro relativo de la emulación de cerebro completo, las IAs neuromórficas y sintéticas, pero ya podemos anunciar otro importante acoplamiento tecnológico: el que existe entre la emulación de cerebro completo y la IA. Incluso si un impulso hacia la emulación de cerebro completo en realidad resulta en emulación de cerebro completo (y no en IA neuromórfica), y aunque la llegada de la emulación de cerebro completo pudiera ser manejada con seguridad, un riesgo aún persistiría: el riesgo asociado a una segunda transición, una transición desde la emulación de cerebro completo hasta la IA, que es una forma en última instancia más poderosa de inteligencia artificial.

Hay muchos otros acoplamientos tecnológicos, que podrían ser considerados en un análisis más exhaustivo. Por ejemplo, un impulso hacia la emulación de cerebro completo impulsaría el progreso neurocientífico en términos generales.[13] Esto podría producir diversos efectos, como un progreso más rápido hacia la detección de mentiras, las técnicas de manipulación neuropsicológicas, la mejora cognitiva, y avances médicos variados. Del mismo modo, un empuje hacia la mejora cognitiva podría (dependiendo de la ruta específica seguida) crear efectos secundarios tales como un desarrollo más rápido de la selección genética y de los métodos de ingeniería genética no sólo con vistas a mejorar la cognición, sino para modificar otras características también.

Anticipándose a las consecuencias

Nos encontramos con otra capa de complejidad estratégica si se tiene en cuenta que no hay ningún controlador del mundo perfectamente benevolente, racional y unificado que simplemente implemente lo que se haya descubierto como la mejor opción. Cualquier punto abstracto

sobre “lo que se debe hacer” debe ser transformado en un mensaje concreto, que se introduce en el ámbito de la realidad retórica y política. Allí será ignorado, mal entendido, distorsionado, o secuestrado para diversos fines conflictivos; rebotará como en un pinball, causando acciones y reacciones, provocando una cascada de consecuencias, el resultado de las cuales no tendrán necesariamente ninguna relación directa con las intenciones del emisor original.

Un operador sofisticado podría tratar de anticipar este tipo de efectos. Consideremos, por ejemplo, la siguiente plantilla argumentativa a favor de continuar con cierta investigación enfocada en desarrollar una tecnología peligrosa X. (Un argumento que encajaría en esta plantilla se puede encontrar en los escritos de Eric Drexler. En el caso de Drexler, X = nanotecnología molecular.[14])

1. Los riesgos de X son grandes.
2. La reducción de estos riesgos requerirá un período de preparación importante.
3. Esta preparación comenzará una vez que amplios sectores de la sociedad tomen en serio las implicaciones de X.
4. Amplios sectores de la sociedad tendrán la posibilidad de tomar X en serlo solamente una vez que un gran esfuerzo de investigación para desarrollar X esté en marcha.
5. Cuanto antes se inicie el esfuerzo de investigación a fondo, más tiempo tardará en llegar X (porque se partirá de un nivel más bajo de tecnologías posibilitantes pre-existentes).
6. Por lo tanto, cuanto antes se inicie el esfuerzo de investigación a fondo, más largo será el período para prepararse y mejor se podrán reducir los riesgos.
7. Por lo tanto, un esfuerzo de investigación serio hacia X debe iniciarse inmediatamente.

Lo que inicialmente parece un motivo para ir lento o detenerse —los riesgos de X son grandes— termina, en esta línea de argumentación, convirtiéndose en un motivo para la conclusión opuesta.

Un tipo relacionado de argumento es que deberíamos —de manera despiadada— aceptar catástrofes pequeñas y medianas porque nos recuerdan nuestras vulnerabilidades y nos impulsan a tomar precauciones que reducen la probabilidad de una catástrofe existencial. La idea es que una catástrofe pequeña o mediana actúa como una vacuna, desafiando a la civilización con una amenaza relativamente superable y estimulando una respuesta inmune que prepare al mundo para hacer frente a la amenaza de variedad existencial.

Estos argumentos de “terapia de shock” abogan por dejar que un mal suceda con la esperanza de que impulse una reacción pública. Los mencionamos aquí no para defenderlos, sino como una manera de introducir la idea de (lo que vamos a llamar) “argumentos de anticipación de consecuencias”. Tales argumentos sostienen que tratando a los demás como seres irracionales y apoyándose en sus

prejuicios y conceptos erróneos es posible obtener una respuesta de ellos más competente que si la cuestión se les hubiera presentado honesta y directamente a sus facultades racionales.

Puede parecer enormemente difícil usar el tipo de estratagemas recomendados por los argumentos de anticipación de consecuencias para alcanzar objetivos globales a largo plazo. ¿Cómo podría alguien predecir el curso final de un mensaje después de haber rebotado de aquí para allá en la máquina de pinball del discurso público? Hacerlo exigiría predecir los efectos retóricos sobre miles de votantes con variadas idiosincrasias y niveles fluctuantes de influencia durante largos períodos de tiempo durante los cuales el sistema podría ser perturbado por eventos no anticipados desde el exterior, mientras que su topología también experimentaría una reorganización endógena continua: ¡sin duda una tarea imposible![16] Sin embargo, puede que no sea necesario hacer predicciones detalladas sobre toda la trayectoria futura del sistema con el fin de identificar una intervención que pudiera esperarse aumentara razonablemente las posibilidades de un determinado resultado a largo plazo. Por ejemplo, podrían considerarse sólo los efectos predecibles y a corto plazo de manera detallada, seleccionando una acción que se ajuste a éstos, mientras se modela el comportamiento del sistema más allá del horizonte previsibilidad como un paseo aleatorio.

Puede haber, sin embargo, argumentos morales para rebajar o abstenerse de argumentos de anticipación de consecuencias. Tratar de ser más listo que el otro parece un juego o de suma cero —o de suma negativa, cuando se tiene en cuenta el tiempo y la energía que se pierde en la práctica, así como la probabilidad de que en general hiciera más difícil para cualquier persona descubrir lo que otros realmente piensan y confiar en ellos al expresar sus propias opiniones.[17] Un despliegue completo de las prácticas de la comunicación estratégica destruiría la sinceridad y dejaría a la verdad incapaz de defenderse de las puñaladas por la espalda de los matones políticos.

Caminos y posibilitadores

¿Deberíamos celebrar los avances de hardware de los ordenadores? ¿Y los avances en el camino hacia la emulación de cerebro completo? Abordaremos estas dos cuestiones a continuación.

Efectos de los avances de hardware

Ordenadores más rápidos hacen que sea más fácil crear inteligencia artificial. Un efecto de acelerar los avances en hardware, por lo tanto,

es acelerar la llegada de la inteligencia artificial. Como se señaló anteriormente, esto es probablemente algo negativo desde una perspectiva impersonal, ya que reduce la cantidad de tiempo disponible para resolver el problema de control y para que la humanidad llegue a una etapa más madura de civilización. El tema no es pan comido, sin embargo. Puesto que la superinteligencia eliminaría muchos otros riesgos existenciales, podría haber razones para preferir un desarrollo temprano si el nivel de estos otros riesgos existenciales fuera muy alto.[18]

Acelerar o retrasar la aparición de la explosión de inteligencia no es el único canal por el cual el ritmo de avance del hardware puede conllevar un riesgo existencial. Otro canal es que el hardware pudiera, en cierta medida, sustituir al software; de manera que, un mejor hardware redujera la habilidad mínima requerida para codificar una IA seminal. Los ordenadores rápidos también podrían fomentar el uso de los enfoques basados en técnicas de fuerza bruta (como los algoritmos genéticos y otros métodos de generación-evaluación-descarte) y menos en las técnicas que requieran una comprensión profunda para su uso. Si las técnicas de fuerza bruta se prestan a diseños de sistema más anárquicos o imprecisos, en los que el problema de control es más difícil de resolver que en los sistemas de ingeniería controlados de manera más precisa y teórica, esto constituiría otra manera en la que los ordenadores más rápidos aumentarían el riesgo existencial.

Otra consideración es que el progreso de hardware rápido aumenta la probabilidad de un despegue rápido. Cuanto más rápidamente avance el estado de la técnica en la industria de semiconductores, menos tiempo personal de los programadores se empleará en la explotación de las capacidades de los ordenadores de cualquier nivel de rendimiento. Esto significa que es menos probable que la explosión de inteligencia se inicie en el nivel más bajo de rendimiento de hardware en el que es factible. Una explosión de inteligencia es, por lo tanto, más probable que se inicie cuando el hardware haya avanzado significativamente más allá del nivel mínimo en el que el enfoque de programación eventualmente exitoso podría haber tenido éxito. Habrá, por lo tanto, un excedente de hardware cuando, finalmente, se produzca el despegue. Como vimos en el capítulo 4, el excedente de hardware es uno de los principales factores que reducen la resistencia al progreso durante el despegue. Los avances en la velocidad del hardware tenderán, por lo tanto, a hacer la transición a la superinteligencia más rápida y explosiva.

Un despegue más rápido gracias a un excedente de hardware puede afectar a los riesgos de la transición de varias maneras. La más obvia es que un despegue más rápido ofrece menos oportunidad de

responder y hacer los ajustes mientras la transición está en marcha, lo que tendería a aumentar el riesgo. Una consideración relacionada es que un excedente de hardware reduciría las posibilidades de que una IA seminal auto-mejorativa peligrosa pudiera ser contenida mediante la limitación de su capacidad de colonizar un hardware suficiente: cuanto más rápido sea cada procesador, menos procesadores necesitará la IA para impulsarse a sí misma hasta la superinteligencia. Sin embargo, otro efecto del excedente de hardware es nivelar el campo de juego entre los grandes y pequeños proyectos mediante la reducción de la importancia de una de las ventajas de los proyectos más grandes —la capacidad de comprar los ordenadores más potentes. Este efecto también podría aumentar el riesgo existencial, si los proyectos más grandes fueran más propensos a resolver el problema de control y a seguir objetivos moralmente aceptables.[19]

También hay ventajas en un despegue más rápido. Un despegue más rápido aumentaría la probabilidad de que se formara una Unidad. Si establecer una Unidad es suficientemente importante para resolver los problemas de coordinación post-transi- ción, podría valer la pena aceptar un mayor riesgo durante la explosión de inteligencia con el fin de mitigar el riesgo de catastróficos fallos de coordinación posteriores.

Los avances en computación pueden afectar el resultado de una revolución de inteligencia artificial no sólo por jugar un papel directo en la construcción de la inteligencia artificial, sino también por tener efectos difusos en la sociedad que indirectamente ayuda a dar forma a las condiciones iniciales de la explosión de inteligencia. Internet, que requirió que el hardware fuera lo suficientemente bueno como para permitir que los ordenadores personales fueran producidos en masa a bajo coste, ahora está influyendo en la actividad humana en muchas áreas, incluyendo el trabajo en inteligencia artificial y la investigación sobre el problema de control. (Este libro podría no haber sido escrito, y es posible que Ud. no lo hubiera encontrado, sin internet). Sin embargo, el hardware ya es lo suficientemente bueno para un gran número de aplicaciones que podrían facilitar la comunicación humana y la deliberación, y no está claro que el ritmo de progreso en estas áreas está fuertemente influenciado por el ritmo de mejoramiento del hardware.[20]

A fin de cuentas, parece que el progreso más rápido en hardware de computación no es deseable desde el punto de vista valorativo impersonal. Esta conclusión tentativa podría ser desechada, por ejemplo, si las amenazas de otros riesgos existenciales o de los fallos de coordinación post-transición resultan ser extremadamente grandes. En cualquier caso, parece difícil tener mucha influencia sobre el ritmo de avance del hardware. Nuestros esfuerzos por mejorar las

condiciones iniciales para la explosión de inteligencia probablemente deben, por lo tanto, centrarse en otros parámetros.

Téngase en cuenta que incluso cuando no podemos ver cómo influir en algún parámetro, puede ser útil determinar su "signo" (es decir, si un aumento o disminución de ese parámetro sería deseable) como un paso preliminar en la cartografía del estado estratégico de la cuestión. Después podríamos descubrir un nuevo punto de apoyo que nos permitiera manipular el parámetro con mayor facilidad. O podríamos descubrir que el signo del parámetro se correlaciona con el signo de algún otro parámetro más manipulable, por lo que nuestro análisis inicial nos ayudaría a decidir qué hacer con este otro parámetro.

¿Debería promoverse la investigación sobre la emulación de cerebro completo?

Cuanto más difícil parezca resolver el problema de control para la inteligencia artificial, más tentador será promover la ruta de emulación de cerebro completo como alternativa menos arriesgada. Hay varias cuestiones, sin embargo, que deben ser analizadas antes de poder llegar a una conclusión bien fundamentada.[21]

En primer lugar está el tema del acoplamiento tecnológico, que se discutió anteriormente. Entonces indicamos que el esfuerzo por desarrollar la emulación de cerebro completo podría acabar dando como resultado una IA neuromórfica, una forma de inteligencia artificial que puede ser especialmente peligrosa.

Pero supongamos, por mor del argumento, que efectivamente lográramos la emulación de cerebro completo (ECC). ¿Sería más segura que la IA? Éste es un tema complicado en sí mismo. Hay por lo menos tres ventajas *putativas* de la ECC: (i) que sus características de rendimiento se entenderían mejor que las de la IA; (ii) que heredaría motivaciones humanas; y (iii) que traería consigo un despegue más lento. Consideremos muy brevemente cada una.

¡ Suena plausible pensar que debería ser más fácil entender las características de rendimiento Intelectual de una emulación que las de una IA. Tenemos abundante experiencia con las fortalezas y debilidades de la inteligencia humana, pero no tenemos experiencia con una inteligencia artificial de nivel humano. Sin embargo, entender lo que un supuesto intelecto humano digitalizado puede y no puede hacer no es lo mismo que entender cómo un intelecto responderá a las modificaciones encaminadas a mejorar su rendimiento. Un intelecto artificial, por el contrario, podría ser cuidadosamente diseñado para ser comprensible, tanto en sus disposiciones estáticas como dinámicas. Así, mientras que la emulación de cerebro completo puede ser más predecible en su rendimiento intelectual que una IA genérica en una etapa de desarrollo comparable, no está claro

si la emulación de cerebro completo sería dinámicamente más predecible que una IA diseñada por programadores competentes y conscientes de la importancia de la seguridad.

ii En cuanto a que una emulación heredaría las motivaciones de su plantilla humana, esto está lejos de ser seguro. La captura de disposiciones evaluativas humanas podría requerir una emulación de muy alta fidelidad. Incluso si las motivaciones de algunos individuos estuvieran perfectamente capturadas, no está claro cuánta seguridad se alcanzaría con ello. Los seres humanos pueden ser poco fiables, egoístas y crueles. Aunque las plantillas con suerte serían seleccionadas por su virtud excepcional, podría ser difícil de predecir cómo actuaría alguien si fuera trasplantado a circunstancias radicalmente extrañas, sobrehumanamente mejorado en inteligencia, y tentado con la oportunidad de dominar el mundo. Es cierto que al menos las emulaciones serían más propensas a tener motivaciones similares a las humanas (en oposición a valorar únicamente la producción de clips o el descubrimiento de los dígitos de pi). Dependiendo de la opinión que tengamos sobre la naturaleza humana, esto podría ser o no ser reconfortante.[22]

iii No está claro por qué la emulación de cerebro completo debería tener como resultado un despegue más lento que la inteligencia artificial. Tal vez se puede esperar menos excedente de hardware con la emulación de cerebro completo, ya que la emulación de cerebro completo sería menos computacionalmente eficiente de lo que podría ser una inteligencia artificial. Tal vez, también, un sistema de inteligencia artificial podría absorber más fácilmente toda la potencia de cálculo disponible en una inteligencia integrada gigante, mientras que la emulación de cerebro completo renunciaría a una superinteligencia de calidad y se destacaría respecto de la humanidad sólo por su velocidad y el tamaño de su población. Si la emulación de cerebro completo condujera a un despegue más lento, esto podría ser beneficioso

iv a la hora de aliviar el problema de control. Un despegue más lento también conllevaría que un resultado multipolar fuera más probable. Pero que un resultado multipolar sea deseable es muy dudoso.

Hay otra complicación importante con la idea general de que conseguir la emulación de cerebro completo primero sea más seguro: la necesidad de hacer frente a una segunda transición. Incluso si la primera forma de inteligencia artificial de nivel humano está basada en la emulación, todavía seguiría siendo viable que se desarrollara la inteligencia artificial. La IA en su forma madura tiene ventajas importantes sobre la ECC, convirtiendo a la IA en la forma de tecnología más poderosa, en última instan- cia.[23] Mientras que una IA madura dejaría a la ECC obsoleta (excepto con el propósito especial de preservar mentes humanas individuales), lo contrario no ocurriría.

Lo que esto significa es que si la IA se desarrollara primero, puede que sólo llegara a haber una ola de explosión de inteligencia. Pero si la ECC se desarrollara primero, podría haber dos oleadas: primero, la llegada de la ECC; y más tarde, la llegada de la IA. El riesgo existencial total a lo largo de la ECC como primer camino es la *suma* de los riesgos de la primera transición y los de la segunda transición (con la condición de haber sobrevivido a la primera); véase la figura 13.[24]

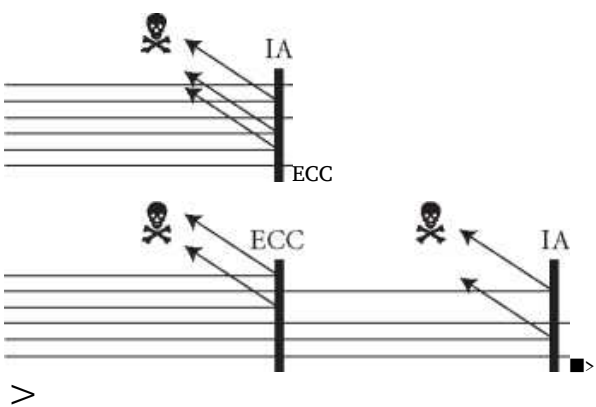


Figura 13. ¿Inteligencia artificial o emulación de cerebro completo primero? En un escenario en que se desarrollara primero la IA, habría una transición que crearía un riesgo existencial. En un escenario en que se desarrolla primero de la ECC, hay dos transiciones de riesgo, primero el desarrollo de la ECC y luego el desarrollo de la IA. El riesgo existencial total a lo largo del escenario en que se desarrollara primero de la ECC es la suma de aquéllos. Sin embargo, el riesgo de una transición de IA podría ser menor si se produjera en un mundo donde la ECC ya se hubiera implantado con éxito.

¿Sería más segura la transición a la IA en un mundo con ECC? Una consideración es que la transición a la IA sería menos explosiva si se

produjera después de que una cierta forma de inteligencia artificial hubiera sido ya alcanzada. Las emulaciones, funcionando a velocidades digitales y en números que podrían exceder con mucho la población humana biológica, reducirían el diferencial cognitivo, por lo que sería más fácil para las emulaciones controlar la IA. Esta consideración no es demasiado consistente, ya que la brecha entre la IA y la ECC podría seguir siendo muy grande. Sin embargo, si las emulaciones fueran no sólo más rápidas y más numerosas, sino también cualitativamente más inteligentes que los humanos biológicos (o al menos estuvieran situadas en el extremo superior de la distribución humana), el escenario de la ECC

como primer camino tendría ventajas paralelas a las del mejoramiento cognitivo humano que comentamos anteriormente.

Otra consideración es que la transición a la ECC ampliaría la ventaja del proyecto puntero. Consideremos un escenario en el que el proyecto puntero tuviera una ventaja de seis meses respecto del seguidor más cercano que estuviera desarrollando la tecnología de emulación de cerebro completo. Supongamos que las primeras emulaciones que se crearan fueran cooperativas, seguras y pacientes. Si se ejecutaran en hardware rápido, estas emulaciones podrían pasar eones subjetivos ponderando cómo crear una IA segura. Por ejemplo, si funcionaran a una velocidad aumentada por 100.000 y fueran capaces de trabajar en el problema de control sin molestias durante seis meses de tiempo sideral, podían trabajar en el problema de control por cincuenta milenios antes de enfrentarse a la competencia de otras emulaciones. Contando con hardware suficiente, podrían acelerar su progreso produciendo miles de copias que trabajaran de forma independiente en subproblemas. Si el proyecto puntero utilizara su ventaja de seis meses para formar una Unidad, podría de este modo dar a su equipo de emulaciones desarrolladoras de IA una cantidad ilimitada de tiempo para trabajar en el problema de control.[25]

A fin de cuentas, parece que el riesgo de la transición a la IA se reduciría si la ECC llegara antes que la IA. Sin embargo, cuando combinamos el riesgo residual de la transición a la IA con el riesgo de la transición a la ECC antecedente, se hace muy claro cómo el riesgo existencial total a lo largo del camino que empieza por la ECC acaba comportando más riesgos que el camino que empieza por la IA. Sólo si uno es muy pesimista sobre la capacidad de la humanidad biológica para gestionar una transición a la IA —teniendo en cuenta que la naturaleza humana o la civilización podrían haber mejorado para cuando se enfrente a ese desafío— podría parecer atractivo el camino que comience por la ECC.

Para averiguar si la tecnología de emulación de cerebro completo debiera ser promovida, tenemos que poner otras cuestiones en la balanza. De manera más significativa, tenemos el acoplamiento tecnológico que se mencionó anteriormente: un impulso hacia la ECC podría producir una IA neuromórfica. Esta es una razón contra la promoción de la ECC.[26] Sin duda, *algunos* diseños sintéticos de IA serían menos seguros que algunos diseños neuromórficos. Sin embargo, sería esperable que los diseños neuromórficos fueran menos seguros. Un motivo para esto es que la imitación puede sustituir a la comprensión. Para construir algo desde cero por lo general hay que tener una comprensión razonablemente buena de cómo funciona el sistema. Tal comprensión puede que no sea necesaria para simplemente copiar características de un sistema existente. La

emulación de cerebro completo se basa en la copia al por mayor de la biología, que puede no requerir una comprensión global a nivel de sistema computacional de la cognición (aunque, sin duda, se necesitaría una gran cantidad de conocimiento a nivel de componentes). La IA neuromórfica puede ser como la emulación de cerebro completo en el siguiente sentido: en que se lograría juntando de manera improvisada piezas plagiadas de la biología sin que los ingenieros tuvieran necesariamente una profunda comprensión matemática de cómo funciona el sistema. Pero la IA neuromórfica sería *diferente* de la emulación de cerebro completo en otro sentido: no tendría motivaciones humanas por defecto.[27] Esta consideración va en contra del enfoque de emulación de cerebro completo en la medida en que probablemente produciría IAs neuromórficas.

Un segundo punto a poner en la balanza es que es más probable que la ECC nos avise de su llegada. Con la IA siempre es posible que alguien haga un avance conceptual inesperado. La ECC, por el contrario, necesitará muchos pasos precursores laboriosos — instalaciones de escaneo de alto rendimiento, software de procesamiento de imágenes, trabajo de modelado neuronal detallado. Por tanto, podemos estar seguros de que la ECC no es inminente (no llegará en menos de, digamos, quince o veinte años). Esto significa que los esfuerzos por acelerar la ECC sólo marcarían la diferencia en escenarios en los que la inteligencia artificial se desarrollara relativamente tarde. Esto podría hacer que las inversiones en ECC fueran atractivas para alguien que quisiera producir una explosión de inteligencia para adelantarse a otros riesgos existenciales, pero no se fiara de apoyar a IA por temor a desencadenar una explosión de inteligencia antes de tiempo, antes de que el problema de control hubiera sido resuelto. Sin embargo, la incertidumbre sobre los plazos relevantes es, probablemente, demasiado grande en la actualidad como para dar a esta consideración mucho peso.[28]

Promocionar la ECC es, por tanto, más atractivo si (a) se es muy pesimista acerca de la capacidad de los seres humanos para resolver el problema de control de la IA, (b) no se está demasiado preocupado por las IAs neuromórficas, los resultados multi- polares o los riesgos de una segunda transición, (c) se cree que el momento esperable para que surjan la ECC y la IA es similar, y (d) se prefiere que la superinteligencia no se desarrolle ni muy tarde ni muy temprano.

La perspectiva de la persona afectada favorece la aceleración

Me temo que el típico e impertinente comentarista de blog pueda

hablar en nombre de muchos cuando él o ella escriba:

> Instintivamente abogo por acelerarlo todo. No porque crea que esto sea lo mejor para el mundo. ¿Por qué debería importarme el mundo cuando esté muerto y enterrado? ¡Quiero acelerarlo, maldita sea! Esto aumentaría mis oportunidades de experimentar un futuro tecnológicamente más avanzado.[29]

Desde el punto de vista de la persona afectada, tenemos una mayor razón para apresurarnos hacia adelante con todo tipo de tecnologías radicales que podrían representar riesgos existenciales. Esto se debe a que el resultado esperable es que casi todos los que ahora existen estén muertos dentro de un siglo.

Los argumentos para apresurarnos son especialmente fuertes en lo que respecta a las tecnologías que podrían extender nuestras vidas y con ello aumentar la fracción esperada de población actualmente existente que todavía podría estar presente cuando acontezca la explosión de inteligencia. Si la revolución de inteligencia artificial fuera bien, la superinteligencia resultante seguramente podría idear medios para prolongar indefinidamente la vida de los seres humanos entonces existentes, no sólo manteniéndolos con vida, sino restaurándoles la salud y el vigor juvenil, y mejorando sus capacidades más allá de lo que actualmente consideramos como las capacidades humanas; o ayudarles a desechar de una vez por todas sus soportes mortales descargando sus mentes en un sustrato digital y dotando a sus espíritus liberados con encarnaciones virtuales llenas de bienestar. En cuanto a las tecnologías que no prometen salvar vidas, el argumento para apresurarse es más débil, aunque quizás tenga suficiente apoyo por la esperanza de elevar las condiciones de vida.[30]

La misma línea de razonamiento hace que la perspectiva de la persona afectada favorezca muchas innovaciones tecnológicas arriesgadas que prometen acelerar el inicio de la explosión de inteligencia, incluso cuando esas innovaciones fueran desfavorecidas desde una perspectiva impersonal. Tales innovaciones podrían acortar las horas peligrosas durante las cuales deberíamos aferrarnos a nosotros mismos para llegar a ver el amanecer de la era posthumana. Desde el punto de vista de la persona afectada, el progreso hacia hardware más rápido parece, por tanto, deseable; al igual que el progreso más rápido hacia la ECC. Cualquier efecto adverso relativo al riesgo existencial probablemente se vea compensado por el beneficio personal de una mayor probabilidad de que la explosión de inteligencia suceda en la vida de las personas que actualmente existen.[31]

Colaboración

Un parámetro importante es el grado en que el mundo conseguirá coordinarse y colaborar en el desarrollo de la inteligencia artificial. La colaboración traería muchos beneficios. Echemos un vistazo a cómo este parámetro podría afectar el resultado y qué mecanismos podríamos tener para aumentar el alcance y la intensidad de la colaboración.

La dinámica de carrera y sus peligros

Existe una dinámica de carrera cuando un proyecto teme estar siendo superado por otro. Esto no requiere la existencia real de múltiples proyectos. Una situación en la que sólo hubiera un proyecto podría exhibir una dinámica de carrera si ese proyecto no fuera consciente de su falta de competidores. Los aliados probablemente no habrían desarrollado la bomba atómica tan rápido como lo hicieron si no hubieran creído (erróneamente) que los alemanes podían estar cerca de la misma meta.

La gravedad de una dinámica de carrera (es decir, el grado en que los competidores prioricen la velocidad sobre seguridad) depende de varios factores, como lo ajustada que esté la carrera, la importancia relativa de la capacidad y de la suerte, el número de competidores, que los equipos que compiten utilicen diferentes enfoques, y el grado en que los proyectos compartan los mismos objetivos. Las creencias de los competidores acerca de estos factores son también relevantes. (Véase cuadro 13).

En el desarrollo de la superinteligencia artificial, parece probable que al menos haya una dinámica de carrera suave, siendo posible que hubiera una dinámica de carrera severa. La dinámica de carrera tiene consecuencias importantes en la forma en que debemos pensar el desafío estratégico que supone la posibilidad de una explosión de inteligencia.

La dinámica de carrera podría impulsar a los proyectos para que avanzaran más rápido hacia la superinteligencia, al tiempo que reduciría la inversión en buscar la solución al problema de control. También son posibles otros efectos perjudiciales adicionales provenientes de la dinámica de carrera, como las hostilidades directas entre competidores. Supongamos que dos naciones están compitiendo por desarrollar la primera superinteligencia, y que uno de ellos parece adelantarse. En una dinámica del “ganador se lo lleva todo”, un proyecto retrasado podría tener la tentación de lanzar

Cuadro 13. *Continúa* entre 1 (resultando en una IA perfectamente segura) y 0 (IA completamente insegura). El eje-x representa la importancia relativa de la capacidad en comparación con la inversión en seguridad a la hora de determinar la velocidad de progreso de un equipo hacia la IA. (En 0,5, el nivel de inversión en seguridad es el doble de importante que la capacidad; en 1, los dos son iguales; en 2, la capacidad es dos veces más importante que el nivel de seguridad; etc). El eje-y representa el nivel de riesgo de la IA (la fracción de utilidad máxima que se espera que consiga el ganador de la carrera).

Vemos que, en todos los escenarios, la peligrosidad de la IA resultante es máxima cuando la capacidad no juega ningún papel, disminuyendo gradualmente a medida que la capacidad crece en importancia.

Objetivos compatibles

Otra forma de reducir el riesgo es dando a los equipos una parte más grande del éxito de los demás. Si los competidores están convencidos de que llegar segundo significa la pérdida total de todo lo que les importa, tomarán cualquier riesgo necesario para superar a sus rivales. Por el contrario, los equipos invertirán más en seguridad si ganar la carrera no es tan decisivo. Esto sugiere que deberíamos fomentar diversas formas de inversiones cruzadas.

El número de competidores

Cuanto mayor es el número de equipos participantes, más peligrosa se vuelve la carrera: cada equipo, al tener menos posibilidades de llegar primero, está más dispuesto a lanzar la precaución por los aires. Esto puede verse al contrastar la Figura 14a (dos equipos) con la Figura 14b (cinco equipos). En todos los escenarios, más competidores implican más riesgo. El riesgo se reduciría si los equipos se unieran en un menor número de coaliciones competidoras.

La maldición de poseer demasiada información

¿Es bueno que los equipos estén al tanto de sus posiciones en la carrera (conociendo sus puntuaciones en capacidad, por ejemplo)? En este caso, existen factores en oposición. Es deseable que un líder sepa que está liderando (para que sepa que tiene algún margen para medidas de seguridad adicionales). Sin embargo, no es deseable que un rezagado sepa que ha quedado a la zaga (ya que esto confirmaría la necesidad de recortar en seguridad para tener alguna esperanza de alcanzar a los primeros). Si bien intuitivamente podría parecer que esta disyuntiva podría interpretarse de varios modos, los modelos son inequívocos: tener información es (en cuanto a su expectativa) negativo.[33] Las Figuras 14a y 14b exponen tres escenarios: las líneas rectas corresponden a situaciones en las que ningún equipo sabe ninguna de las puntuaciones en capacidad, ni las suyas propias. Las líneas discontinuas muestran situaciones en las que cada equipo sólo conoce su propia capacidad. (En esas situaciones, un equipo tiene riesgo adicional sólo si su

capacidad es baja). Y las líneas de puntos muestran lo que sucede cuando todos los equipos conocen las capacidades de todos. (Asumiendo riesgos adicionales si sus puntuaciones de capacidad son similares a los de los demás). Con cada incremento en el nivel de información, peor se vuelve la dinámica de carrera.

un ataque desesperado contra su rival en lugar de esperar pasivamente la derrota. Previendo esta posibilidad, el proyecto adelantado podría tener la tentación de atacar preventivamente. Si los antagonistas son Estados poderosos, el choque podría ser sangriento.[34] (Un "ataque quirúrgico" contra el proyecto de IA de los rivales podría correr el riesgo de desencadenar una confrontación más grande y podría, en todo caso, no ser factible si el país agredido hubiera tomado precauciones)[35].

Los escenarios en los que los desarrolladores rivales no son Estados, sino entidades más pequeñas, tales como laboratorios corporativos o equipos académicos, probablemente supondrían que habría mucha menos destrucción directa en caso de conflicto. Sin embargo, las consecuencias generales de la competencia pueden ser casi igual de negativas. Esto se debe a que la mayor parte del daño esperado por la competencia no se deriva del violento choque de la batalla, sino de la rebaja en precaución. Una dinámica de carrera llevaría, como ya vimos, a reducir la inversión en seguridad; y el conflicto, aunque no fuera violento, tendería a eliminar las oportunidades de colaboración, ya que los proyectos serían menos propensos a compartir ideas para resolver el problema de control en un clima de hostilidad y desconfianza.[36]

Sobre los beneficios de la colaboración

Por tanto, la colaboración ofrece muchos beneficios. Reduce el apresuramiento por desarrollar inteligencia artificial. Permite una mayor inversión en seguridad. Evita los conflictos violentos. Y facilita el intercambio de ideas sobre cómo resolver el problema de control. A estos beneficios podemos añadir otro: la colaboración tendería a producir resultados en los que los frutos de una explosión de inteligencia controlada con éxito se distribuirían de manera más equitativa.

Que una colaboración más amplia debiera traducirse en un mayor reparto de las ganancias no es un axioma. En principio, un pequeño proyecto dirigido por un altruista podría conducir a un resultado donde los beneficios se compartieran uniforme o equitativamente entre todos los seres susceptibles de consideración moral. Sin embargo, hay varias razones para suponer que las colaboraciones más amplias, con la participación de un mayor número de patrocinadores, serán (esperamos) superiores en un sentido distributivo. Una de esas

razones es que los patrocinadores presumiblemente preferirán un resultado en el que ellos mismos recibieran (por lo menos) lo que les corresponda. Una amplia colaboración implica entonces que muchos individuos conseguirán al menos su justa parte, asumiendo que el proyecto tuviera éxito. Otra razón es que también parece más probable que una amplia colaboración beneficiara a personas fuera de la colaboración. Una colaboración más amplia tendría más miembros, de manera que más personas ajenas al proyecto tendrían vínculos personales con alguien de dentro que miraría por sus intereses. También es más probable que una colaboración más amplia incluya al menos a algunos altruistas que quieran beneficiar a todos. Por otra parte, es más probable que una colaboración más amplia operara bajo supervisión pública, lo que podría reducir el riesgo de que todo el pastel fuera aglutinado por un conjunto de programadores o inversores.[37] Nótese también que cuanto mayor sea la colaboración exitosa, menor es el coste de extenderla a todos los forasteros. (Por ejemplo, si el 90% de todas las personas estuvieran dentro de la colaboración, no les costaría más del 10% de sus propiedades poner a todos los forasteros a su mismo nivel).

Por tanto, es plausible que colaboraciones más amplias lleven a una distribución más amplia de los beneficios (aunque *algunos* proyectos con pocos patrocinadores también pueden tener objetivos distributivos excelentes). Pero ¿por qué es deseable una gran distribución de ganancias?

Hay razones tanto morales como prudenciales para favorecer resultados en los que todo el mundo tenga una parte de la recompensa. No vamos a decir mucho sobre el argumento moral, salvo para señalar que no necesita descansar en ningún principio igualitario. El argumento podría basarse, por ejemplo, en razones de equidad. Un proyecto que creara la superinteligencia artificial impondría un riesgo externo global. Todo el mundo en el planeta se pone en peligro, incluyendo a aquellos que no den su consentimiento a que sus propias vidas y las de sus familias sean puestas en peligro de esta manera. Ya que todo el mundo comparte el riesgo, parecería un requisito mínimo de equidad que todo el mundo obtuviera también una parte del beneficio.

El hecho de que la cantidad total (esperada) de bien parece mayor en los escenarios de colaboración es otra razón importante para que tales escenarios sean moralmente preferibles.

El argumento prudencial a favor de una amplia distribución de las ganancias tiene dos vertientes. Una vertiente consiste en que una distribución amplia debería promover la colaboración, mitigando así las consecuencias negativas de la dinámica de carrera. Hay menos incentivos para luchar por construir la primera superinteligencia si

todo el mundo se beneficia por igual del éxito de cualquier proyecto. Los patrocinadores de un proyecto en particular también podrían beneficiarse de señalar de manera creíble su compromiso de distribuir las ganancias universalmente, un proyecto altruista certificable probablemente atraería a más seguidores y a menos enemigos.[38]

La otra vertiente del argumento prudencial a favor de una amplia distribución de las ganancias tiene que ver con que los agentes tengan aversión al riesgo o tengan funciones de utilidad que fueran sublineales en cuanto a recursos. El hecho central en este caso es la enormidad de la potencial ganancia de recursos. Suponiendo que el universo observable esté tan deshabitado como parece, contiene más de una galaxia vacante por cada ser humano vivo. La mayoría de la gente preferiría tener acceso a los recursos de una galaxia que a un billete de lotería que ofreciera una-entre-mil millones de oportunidades de ser dueño de mil millones de galaxias.[39] Dado el tamaño astronómico de los recursos cósmicos de la humanidad, parece que el interés propio generalmente debe favorecer acuerdos que garanticen a cada persona una parte, incluso si cada parte corresponde a una pequeña fracción del total. Lo importante, cuando tenemos la perspectiva de una bonanza tan extravagante, es no quedarse fuera a la intemperie.

Este argumento de la enormidad del pastel de recursos supone que las preferencias son susceptibles de ser satisfechas con recursos.[40] Esta suposición no se cumple necesariamente. Por ejemplo, varias teorías éticas prominentes —incluyendo teorías consecuencialistas especialmente agregativas— se basan en funciones de utilidad de riesgo neutral y lineal en cuanto a recursos. Mil millones de galaxias podrían utilizarse para crear un billón de veces más vidas felices que una sola galaxia. Son, por tanto, para un utilitario, mil millones de veces más valiosas.[41] Funciones de preferencia ordinarias de humanos egoístas, sin embargo, parecen ser relativamente susceptibles de ser saciadas por recursos.

Esta última afirmación debe ser acompañada de dos importantes matizaciones. La primera es que muchas personas se preocupan por el estatus. Si varios agentes quieren encabezar la lista de ricos de Forbes, entonces no habría un pastel de recursos lo suficientemente grande como para que todo el mundo lograra plena satisfacción.

El segundo requisito es que la base de la tecnología de post-transición permitiría convertir los recursos materiales en una gama sin precedentes de productos, incluyendo algunos productos que no están disponibles actualmente a ningún precio a pesar de que son muy apreciados por muchos seres humanos. Un milmillonario no vive mil más años que un millonario. En la era de las mentes digitales, sin embargo, el multimillonario podría permitirse una potencia de cálculo

mil veces mayor y así podría disfrutar de una esperanza de vida subjetiva mil veces más larga. La capacidad mental, del mismo modo, podría estar a la venta. En tales circunstancias, con el capital económico convirtiéndose en bienes vitales a una velocidad constante incluso para grandes niveles de riqueza, la codicia sin límites tendría más sentido del que tiene en el mundo de hoy, donde los ricos (aquellos que carecen de un corazón filantrópico) se dedican a gastar sus riquezas en aviones, barcos, colecciones de arte, o en una cuarta y quinta residencia.

¿Significa esto que un egoísta debería ser neutral respecto al riesgo que podría correr su dotación de recursos después de la transición? No exactamente. Los recursos físicos podrían no ser convertibles en duración o en rendimiento mental a escalas arbitrarias. Si una vida debe ser vivida de forma secuencial, de modo que los momentos de observación puedan recordar eventos anteriores y ser afectados por las decisiones previas, entonces la vida de una mente digital no puede ser extendida de forma arbitraria sin la utilización de un número creciente de operaciones computacionales *secuenciales.* Pero la física limita el grado en que los recursos pueden transformarse en computaciones secuenciales.[42] Los límites de la computación secuencial también podrían restringir algunos aspectos del rendimiento cognitivo a escalar de manera radicalmente sublineal más allá de una dotación de recursos relativamente modesta. Por otra parte, no es obvio que un egoísta fuera o debiera ser neutral ante el riesgo, incluso con respecto a las métricas de resultado altamente relevantes en un sentido normativo, como el número subjetivo de años de vida de calidad. Si se le ofreciera la posibilidad de elegir entre un extra de 2.000 años de vida con seguridad y una probabilidad de uno entre diez de un extra de 30.000 años de vida, creo que la mayoría de la gente elegiría la primera (incluso bajo la condición de que cada año de vida fuera de igual calidad).[43]

En realidad, el argumento prudencial a favor de una amplia distribución de las ganancias es presumiblemente relativo a cada sujeto y dependiente de cada situación. Sin embargo, en general, las personas tendrían más posibilidades de conseguir (casi todo) lo que quisieran si encontraran una manera de lograr una amplia distribución —y esto es así incluso sin tener en cuenta que el compromiso con una distribución más amplia tendería a fomentar la colaboración y de ese modo aumentarían las posibilidades de evitar una catástrofe existencial. Favorecer una distribución amplia, por lo tanto, parece ser no sólo moralmente obligatorio, sino también prudencialmente aconsejable.

Hay otra serie de consecuencias de la colaboración que al menos deben ser mencionadas: la posibilidad de que la colaboración pre-

transición influya en el nivel de colaboración post-transición. Asumamos que la humanidad resuelve el problema de control. (Si el problema de control no se resolviera, apenas importaría cuánta colaboración hubiera post-transición). Hay dos escenarios a considerar. El primero es que la explosión de inteligencia *no* cree una dinámica de "el ganador se lo lleva todo" (probablemente porque el despegue fuera relativamente lento). En este caso, es posible que si la colaboración pre-transición fuera a tener algún efecto sistemático en la colaboración post-transición, tendría un efecto positivo, que tendería a promover la colaboración posterior. Las relaciones de colaboración originales pueden aguantar y continuar más allá de la transición; asimismo, la colaboración pre-transición podría ofrecer más oportunidades a las personas de dirigir la evolución post-transición en direcciones más deseables (y, presumiblemente, más colaborativas).

El segundo escenario es que la naturaleza de la explosión de inteligencia no fomentara una dinámica de "el ganador se lo lleva todo" (probablemente porque el despegue fuera relativamente rápido). En este caso, si no hubiera una amplia colaboración antes del despegue, es probable que surgiera una Unidad —un solo proyecto experimentaría la transición, obteniendo en algún momento una ventaja estratégica decisiva combinada con superinteligencia. Una Unidad, por definición, es un orden social altamente colaborativo.[44] La ausencia de una amplia colaboración pre-transición daría, por lo tanto, lugar a un grado extremo de colaboración post-transición. En contraste, un nivel algo más elevado de colaboración en el período previo a la explosión de inteligencia abre una variedad más amplia de posibles resultados. Proyectos colaborativos podrían sincronizar su ascenso para asegurarse de llevar a cabo la transición simultáneamente, sin que ninguno de ellos consiguiera una ventaja estratégica decisiva. O diferentes grupos patrocinadores podrían fusionar sus esfuerzos en un solo proyecto, mientras renuncian a dar a ese proyecto la orden de formar una Unidad. Por ejemplo, uno podría imaginar un consorcio de naciones formando un proyecto científico conjunto para desarrollar la superinteligencia artificial, pero sin permitir que este proyecto se convirtiera en algo parecido a unas Naciones Unidas sobrecargadas, eligiendo en su lugar mantener el orden mundial faccioso que existía previamente.

Particularmente en el caso de un despegue rápido, existe la posibilidad de que una mayor colaboración pre-transición resultara en menos colaboración post-tran- sición. Sin embargo, en la medida en que las entidades colaboradoras son capaces de influir en el resultado, pueden permitir la aparición o la persistencia de la no-colaboración sólo si prevén que el sectarismo post-transición no tendrá

consecuencias catastróficas. Los escenarios en los que la colaboración pre-transición condujera a una colaboración post-transición reducida pueden ser principalmente aquellos en los que la reducción de la colaboración post-transición fuera inocua.

En general, una mayor colaboración post-transición parece deseable. Se reduciría el riesgo de dinámicas distópicas en las que la competencia económica y una rápida expansión de la población llevaría a una situación malthusiana, o en las que la selección evolutiva erosionara los valores humanos y seleccionara formas no-felicitarias, o en las que las potencias enfrentadas sufrirían otros fallos de coordinación, tales como guerras y carreras tecnológicas. El último de estos temas, la perspectiva de carreras tecnológicas, puede ser especialmente problemática si la transición nos lleva a una forma intermedia de inteligencia artificial (la emulación de cerebro completo), puesto que crearía una nueva dinámica de carrera que mermaría las posibilidades de poder resolver el problema de control para la segunda transición hacia una forma más avanzada de la inteligencia artificial (IA).

Hemos descrito anteriormente cómo la colaboración podría reducir los conflictos en el período previo a la explosión de inteligencia, lo que aumentaría las posibilidades de que el problema de control se resolviera, y mejoraría tanto la legitimidad moral como la conveniencia prudencial de la asignación de recursos resultante. A estos beneficios de la colaboración puede ser posible añadir uno más: que una colaboración pre-transición más amplia podría ayudar con los problemas de coordinación importantes de la era post-transición.

Trabajando juntos

La colaboración puede adoptar diferentes formas dependiendo de la escala de las entidades colaboradoras. A pequeña escala, los equipos individuales de IA que crean estar en competencia entre sí podrían elegir poner en común sus esfuerzos.[45] Las corporaciones podrían fusionar o realizar inversiones cruzadas. En una escala mayor, los Estados podrían unirse en un gran proyecto internacional. Existen precedentes de colaboración internacional a gran escala en ciencia y tecnología (como el CERN, el Proyecto Genoma Humano, y la Estación Espacial Internacional), pero un proyecto internacional para desarrollar una superinteligencia segura plantearía un desafío de diferente dimensión debido a las implicaciones en materia de seguridad. Tendría que estar constituido no como una colaboración académica abierta, sino como una empresa conjunta muy bien controlada. Tal vez los científicos involucrados tendrían que estar

físicamente aislados e impedidos de comunicarse con el resto del mundo durante la duración del proyecto, excepto a través de un único canal de comunicación cuidadosamente examinado. El nivel de seguridad requerido podría ser casi inalcanzable en la actualidad, pero los avances en detección de mentiras y en tecnología de vigilancia podría hacerlo factible para finales de este siglo. También vale la pena tener en cuenta que una amplia colaboración no significa necesariamente que un gran número de investigadores estén involucrados en el proyecto; simplemente significa que muchas personas tendrían algo que decir sobre los objetivos del proyecto. En principio, un proyecto podría implicar una colaboración máximamente amplia que incluyera a toda la humanidad como patrocinadores (representados, digamos, por la Asamblea General de las Naciones Unidas), pero que, sin embargo, empleara a un solo científico para llevar a cabo la tarea.[46]

Hay una razón para iniciar la colaboración lo más pronto posible, principalmente para aprovechar el velo de la ignorancia que nos oculta toda información específica sobre qué proyecto individual llegará primero a la superinteligencia. Cuanto más cerca de la línea de llegada estemos, menos incertidumbre habrá sobre las posibilidades relativas de los proyectos en competencia; y más difícil será, en consecuencia, proponer un argumento que defienda el interés que pudiera tener el proyecto avanzado de unirse a un proyecto de colaboración que distribuyera los beneficios a toda la humanidad. Por otro lado, también parece difícil establecer una colaboración formal de alcance mundial antes de que la superinteligencia se haya convertido en un tema mucho más reconocido de lo que actualmente es, y antes de que haya un camino claramente visible que estuviera llevándonos a la creación de la superinteligencia artificial. Por otra parte, en la medida en que la colaboración promovería el progreso, podría llegar a ser contraproducente en términos de seguridad, como se explicó anteriormente.

La forma ideal de colaboración para el presente, por lo tanto, puede ser una que no requiera inicialmente acuerdos específicamente formalizados y que no acelere los avances en inteligencia artificial. Una propuesta ajustada a estos criterios consistiría en proponer una norma moral adecuada que expresara nuestro compromiso con la idea de que la superinteligencia debería contribuir al bien común. Una norma de este tipo podría ser formulada como sigue:

El principio del bien común

La superinteligencia sólo debe desarrollarse para el beneficio de toda la humanidad y al servicio

de ideales éticos ampliamente compartidos.[47]

Establecer desde el principio que el inmenso potencial de la

superinteligencia pertenece a toda la humanidad dará más tiempo para que dicha norma se afiance.

El principio del bien común no excluye incentivos comerciales para individuos o empresas activas en áreas relacionadas. Por ejemplo, una empresa podría satisfacer la petición de compartir universalmente de los beneficios de la superinteligencia mediante la adopción de una "cláusula de bonanza" que implicaría que los beneficios hasta algún tope muy alto (por ejemplo, un billón de dólares anuales) se distribuirían de manera ordinaria entre los accionistas de la empresa y otros demandantes legales, y que sólo los beneficios que superaran ese umbral serían distribuidos a toda la humanidad de manera uniforme (o de otra manera de acuerdo a criterios morales universales). La adopción de unas "cláusulas de bonanza" de este tipo debería ser prácticamente gratuita, pues sería muy poco probable que ninguna empresa superara dicho umbral de beneficio estratosférico (y tales escenarios de baja probabilidad no juegan habitualmente ningún papel en las decisiones de los gestores e inversores de la empresa). Sin embargo, su adopción generalizada daría a la humanidad una garantía valiosa (en la medida en que se pudiera confiar en tales compromisos) de que si alguna vez alguna empresa privada *diera* en el clavo con la explosión de inteligencia, todo el mundo compartiría la mayor parte de los beneficios. La misma idea se podría aplicar a entidades distintas de las empresas. Por ejemplo, los Estados podrían estar de acuerdo en que si alguna vez el PIB de cualquier Estado superara una fracción muy alta (por ejemplo, el 90%) del PIB mundial, el exceso debería distribuirse de manera uniforme entre todos.[48]

El principio del bien común (y los ejemplos particulares, tales como las "cláusulas de bonanza") podrían adoptarse inicialmente como un compromiso moral voluntario por parte de los individuos y las organizaciones responsables que trabajen en áreas relacionadas con la inteligencia artificial. Más tarde, podría ser aprobado por un conjunto más amplio de entidades y convertirse en ley y tratado. Una formulación vaga, como la que aquí se da, podría servir bien como punto de partida; pero en última instancia, debería ser perfilada en un conjunto de requisitos específicos verificables.

CAPÍTULO 15

La hora de la verdad

N

os encontramos en un bosque de complejidad estratégica, rodeado de una densa niebla de incertidumbre. Aunque se han discutido muchas cuestiones, sus detalles y relaciones siguen siendo poco claras y dudosas, y podría haber otros factores en los que ni siquiera hemos reparado todavía. ¿Qué podemos hacer en esta situación?

Filosofía con fecha límite

A un colega mío le gusta señalar que la medalla Fields (el honor más alto en matemáticas) indica dos cosas sobre el agraciado: que era capaz de lograr algo importante, y que no lo hizo. Aunque es una observación dura, no carece de verdad.

Pensemos en un "descubrimiento" como un acto que trae una determinada información desde un punto posterior en el tiempo a un estado anterior. El valor del descubrimiento no se debe al valor de la información descubierta, sino más bien al valor de tener la información disponible antes de lo esperado. Un científico o un matemático puede mostrar una gran habilidad al ser el primero en encontrar una solución que había eludido a muchos otros; sin embargo, si el problema se hubiera resuelto pronto de todos modos, entonces el trabajo probablemente no beneficiaría mucho al mundo.

Hay casos en los que tener una solución sólo un poco más pronto es inmensamente valioso, pero esto es más plausible cuando la solución se usa inmediatamente, ya sea aplicada a algún fin práctico o sirviendo como base para futuros trabajos teóricos. Y en este último caso, cuando la solución se utiliza inmediatamente sólo como un bloque de construcción para su posterior teorización, hay un gran valor en la obtención de una solución ligeramente más pronto sólo si el trabajo extra que permite es importante y urgente.[1]

La pregunta, entonces, no es si el resultado descubierto por el medallista Fields es "importante" en sí mismo (ya sea instrumentalmente o por el propio bien del conocimiento). La pregunta es más bien si fue importante que el medallista permitiera la publicación del resultado antes de tiempo. El valor de este transporte temporal debe ser comparado con el valor que una mente matemática de primer nivel podría haber generado trabajando en otra cosa. Al menos en algunos casos, la medalla Fields podría indicar una vida dedicada a la solución del problema equivocado —por ejemplo, un problema cuyo encanto consistiera principalmente en ser notablemente difícil de resolver.

Dardos similares podrían lanzarse a otros campos, como la filosofía académica. La filosofía trata algunos problemas que son relevantes para mitigar el riesgo existencial —nos encontramos varios casos en este libro. Sin embargo, hay también subcampos dentro de la filosofía que no tienen relación aparente con los riesgos existenciales o, de hecho, con ninguna preocupación práctica. Al igual que con las matemáticas puras, algunos de los problemas estudiados por la filosofía podrían considerarse como intrínsecamente importantes, en el sentido de que los seres humanos tienen razones para preocuparse por ellos independientemente de cualquier aplicación práctica. Comprender la naturaleza fundamental de la realidad, por ejemplo, podría ser algo importante en sí mismo. El mundo sería sin duda menos glorioso si nadie estudiara metafísica, cosmología, o teoría de cuerdas. Sin embargo, la perspectiva auroral de una explosión de inteligencia arroja nueva luz sobre esta antigua búsqueda de la sabiduría.

La perspectiva sugiere ahora que el progreso filosófico puede ser maximizado a través de una ruta indirecta diferente a filosofar de manera inmediata. Una de las muchas tareas en las que la superinteligencia (o incluso una inteligencia humana moderadamente mejorada) superaría al elenco actual de pensadores es en la tarea de responder a las preguntas fundamentales de la ciencia y la filosofía. Esta reflexión sugiere una estrategia de gratificación diferida. Podríamos posponer el trabajo en algunas de las preguntas eternas por un tiempo, delegando esa tarea a nuestros —esperemos— más

competentes sucesores, con el fin de centrar nuestra atención en un desafío más apremiante: el aumento de probabilidad de que lleguemos a tener realmente sucesores competentes. Esto supondría un gran impacto en filosofía y en matemáticas.[2]

¿Qué debemos hacer?

Por tanto, queremos centrarnos en los problemas que no sólo son importantes sino urgentes en el sentido de que necesitamos sus soluciones antes de la explosión de inteligencia. También debemos estar atentos a no trabajar en problemas que tengan un valor negativo (aquellos cuya solución sea perjudicial). Algunos problemas técnicos en el campo de la inteligencia artificial, por ejemplo, podrían tener valor negativo por cuanto su solución podría acelerar el desarrollo de la inteligencia artificial sin hacer demasiado para acelerar el desarrollo de métodos de control que podrían hacer que sobreviviéramos a la revolución en inteligencia artificial y que fuera algo beneficioso.

Puede ser difícil identificar qué problemas son urgentes e importantes y cuáles podemos afrontar con confianza en que tendrán un valor positivo. La incertidumbre estratégica que rodea la mitigación del riesgo existencial significa que debemos preocuparnos de que las intervenciones, incluso las bien intencionadas, no lleguen a ser no sólo improductivas, sino contraproducentes. Para limitar el riesgo de hacer algo activamente perjudicial o moralmente incorrecto, debemos preferir trabajar en problemas que parezcan *sólidamente de valor positivo* (es decir, cuya solución conllevaría una contribución positiva en una amplia gama de escenarios) y emplear los medios que estén sólidamente justificados (es decir, que sean aceptables desde una amplia gama de puntos de vista morales).

Hay un desiderátum más a considerar a la hora de seleccionar qué problemas prio- rizar. Queremos trabajar en problemas que sean *elásticos* a nuestros esfuerzos de solucionarlos. Los problemas muy elásticos son aquellos que se pueden resolver mucho más rápido, o resolverse en un grado mucho mayor, dada una unidad adicional de esfuerzo. Fomentar más la bondad en el mundo es un problema importante y urgente —uno, además, que parece tener un valor positivo consistente: sin embargo, en ausencia de una idea revolucionaria para saber cómo conseguirlo, probablemente sea un problema muy poco elástico. El logro de la paz mundial, del mismo modo, sería muy deseable; pero teniendo en cuenta los numerosos esfuerzos ya orientados hacia ese problema, y los obstáculos formidables que impiden una solución rápida, parece poco probable que las contribuciones de unos pocos individuos adicionales puedan

marcar una gran diferencia.

Para reducir los riesgos de la revolución de la inteligencia artificial, propondremos dos objetivos que parecen satisfacer mejor todos estos desiderátums: la visión estratégica y la creación de capacidad. Podemos estar relativamente seguros del sentido de estos parámetros —cuanta más visión estratégica y más capacidad, mejor. Además, los parámetros son elásticos: una pequeña inversión extra puede significar una diferencia relativamente grande. Ganar en visión y en capacidad también es urgente porque los primeros refuerzos a estos parámetros pueden acumularse, haciendo que los esfuerzos posteriores sean más eficaces. Además de estos dos grandes objetivos, vamos a señalar algunos otros objetivos potencialmente valiosos para las iniciativas.

Buscar la luz estratégica

En un contexto de perplejidad e incertidumbre, el análisis destaca por tener un valor esperado especialmente alto.[3] La iluminación de nuestra situación estratégica nos ayudaría a orientar las intervenciones posteriores de manera más efectiva. El análisis estratégico es especialmente necesario cuando dudamos radicalmente no sólo sobre algunos detalles periféricos, sino sobre las cualidades cardinales de temas centrales. Para muchos parámetros clave, dudamos radicalmente incluso de su *signo,* es decir, no sabemos qué dirección de cambio sería deseable y cuál indeseable. Nuestra ignorancia podría no ser irremediable. El campo ha sido poco prospectado y todavía podría suceder que hubiera brillantes intuiciones de visión estratégica esperando a ser desenterradas a pocos metros debajo de la superficie.

Lo qué aquí queremos decir con “análisis estratégico” es una búsqueda de consideraciones cruciales: las ideas o argumentos con el potencial de cambiar nuestros puntos de vista, no sólo acerca de la estructura de implementación, sino de la topología general de su deseabilidad.[4] Incluso una sola consideración crucial perdida podría viciar nuestros esfuerzos más valientes o hacerlos tan activamente perjudiciales como los de un soldado que luchara en el bando equivocado. La búsqueda de consideraciones cruciales (que debería explorar cuestiones normativas así como descriptivas) a menudo requerirá entrecruzar las fronteras entre distintas disciplinas académicas y otros campos del conocimiento. Como no existe una metodología establecida sobre cómo hacer este tipo de investigación, el difícil pensamiento original es necesario.

Desarrollar una buena capacidad

Otra actividad de alto valor, que comparte con el análisis estratégico la consistente propiedad de ser beneficiosa en una amplia gama de escenarios, es el desarrollo de una base de apoyo bien constituida que tome en serio el futuro. Una base de este tipo puede proporcionar de inmediato los recursos para la investigación y el análisis. Si otras prioridades se hicieran visibles, los recursos podrían ser redirigidos en consecuencia. Una base de apoyo es, pues, una capacidad de propósito general cuyo uso pudiera ser guiado por las nuevas ideas que fueran surgiendo.

Un activo valioso sería una red de donantes que comprendiera a las personas dedicadas a la filantropía racional, informadas sobre los riesgos existenciales, y conscientes de los medios de mitigación. Sería especialmente deseable que los financia- dores iniciales fueran astutos y altruistas, ya que podrían tener oportunidades para dar forma a la cultura de ese ámbito antes de que los habituales intereses venales tomaran posición y se afianzaran. El enfoque durante estas tácticas de apertura debería centrarse en contratar el tipo correcto de personal para dicho ámbito. Podría valer la pena renunciar a algunos avances técnicos en el corto plazo con el fin de llenar los equipos con personas que realmente se preocuparan por la seguridad y que estuvieran orientados hacia la búsqueda de la verdad (y que probablemente pudieran atraer a más personas similares).

Una variable importante es la calidad de la "epistemología social" del campo de la IA y de sus principales proyectos. Descubrir consideraciones cruciales es valioso, pero sólo si afectan a la acción. Esto no siempre se puede dar por sentado. Imaginemos un proyecto que invirtiera millones de dólares y años de trabajo para desarrollar un prototipo de IA, y que después de superar muchos desafíos técnicos el sistema estuviera finalmente empezando a mostrar un progreso real. Existe la posibilidad de que con un poco más de trabajo pudiera convertirse en algo útil y rentable. En ese momento, se descubre una consideración crucial que indica que un enfoque completamente diferente sería un poco más seguro. ¿Se debería suicidar el proyecto como un samurai deshonrado, renunciando a su diseño inseguro y a todo el progreso que se había logrado? ¿O reaccionaría como un pulpo asustado, hinchando una nube de escepticismo con la esperanza de eludir el ataque? Un proyecto que eligiera de forma fiable la opción del samurai ante tales dilemas sería un desarrollador claramente preferible.[5] Sin embargo, construir procesos e instituciones que estén dispuestas a cometer el seppuku basado en acusaciones inciertas y en razonamientos especulativos, no es fácil. Otra dimensión de la epistemología social es la gestión de información confidencial, en

particular, la capacidad de evitar fugas de información que deberían mantenerse en secreto. (La contención de información puede ser especialmente difícil para los investigadores académicos, acostumbrados como están a difundir constantemente sus resultados en cada farola y árbol disponible).

Medidas particulares

Además de los objetivos generales de la luz estratégica y la buena capacidad, algunos objetivos más específicos también podrían presentar oportunidades rentables para la acción. Uno de ellos es el progreso en los desafíos técnicos de seguridad de la inteligencia artificial. Mientras persigamos este objetivo, deberíamos tener cuidado gestionando la información peligrosa. Parte del trabajo que sería útil para resolver el problema de control también sería útil para resolver el problema de la competencia. Un trabajo que queme el fusible de la IA fácilmente podría ser negativo en última instancia.

Otro objetivo específico es promover las "buenas prácticas" entre los investigadores de la IA. Todo avance sobre el problema de control necesita ser compartido. Algunas formas de experimentación computacional, especialmente si involucran una fuerte auto-mejora recursiva, también podrían requerir el uso de control de capacidades para mitigar el riesgo de un despegue accidental. Mientras que la aplicación efectiva de métodos de seguridad no es tan relevante hoy en día, se volverá cada vez más importante a medida que el avance el estado de la técnica. Y no es demasiado pronto para hacer un llamamiento a los profesionales para que expresen un *compromiso con la seguridad,* incluyendo la firma del principio del bien común y la promesa de redoblar esfuerzos en seguridad cuando la perspectiva de la superinteligencia artificial comience a parecer más inminente. Las palabras piadosas no son suficientes y sólo con ellas no se convertirá a una tecnología peligrosa en algo seguro: pero quizás nuestras mentes acaben yendo donde van nuestras palabras.

Otras oportunidades también podrían surgir de vez en cuando para impulsar algún parámetro fundamental, por ejemplo, para mitigar algún otro riesgo existencial, o para promover la mejora biológica cognitiva y las mejoras de nuestra sabiduría colectiva, o incluso para cambiar la política mundial hacia un registro más armonioso.

Que los mejores en naturaleza humana, por favor, se pongan en pie

Ante la perspectiva de una explosión de inteligencia, los humanos somos como niños pequeños jugando con una bomba. Tal es el desajuste entre el poder de nuestro juguete y la inmadurez de nuestra conducta. La superinteligencia es un reto para el que no estamos listos ahora y para el que no estaremos preparados en un largo tiempo. Tenemos poca idea de cuándo se producirá la detonación, aunque si mantenemos el dispositivo cerca de nuestro oído podemos escuchar un débil sonido de tic-tac.

Para un niño con una bomba a punto de detonar en sus manos, algo sensato sería dejarla con cuidado en el suelo, salir rápidamente de la habitación, y ponerse en contacto con el adulto más cercano. Sin embargo, lo que tenemos aquí no es un niño, sino muchos, cada uno con acceso a un gatillo de disparo independiente. Las posibilidades de que *todos* tengamos el juicio suficiente como para abandonar los artilugios peligrosos son casi insignificantes. Algún pequeño idiota inevitablemente pulsará el botón de encendido sólo para ver qué pasa.

Tampoco podemos encontrar un lugar seguro huyendo, pues la detonación de una explosión de inteligencia haría que el firmamento se derrumbara por completo. Tampoco hay ningún adulto a la vista.

En esta situación, cualquier sentimiento de euforia iluminada estaría fuera de lugar. La consternación y el miedo estarían más cerca de acertar; pero la actitud más adecuada podría ser una determinación implacable en ser tan competente como podamos, tanto como si nos estuviéramos preparando para un examen difícil que, o bien nos llevaría a cumplir nuestros sueños, o los destruiría para siempre.

Esto no es una prescripción para el fanatismo. La explosión de inteligencia todavía podría tardar muchas décadas en llegar. Además, nos enfrentamos también al desafío de aferrarnos a nuestra humanidad: de mantener nuestras raíces, sentido común y jovial decencia incluso en las fauces del problema más antinatural e inhumano. Tenemos que poner todo nuestro ingenio humano a trabajar en su solución.

Sin embargo, no debemos perder de vista qué es universalmente importante. Más allá de la niebla de trivialidades cotidianas, podemos percibir —aunque sea débilmente— la tarea esencial de nuestra época. En este libro, hemos tratado de discernir los rasgos de lo que no deja de ser una visión relativamente amorfa y negativamente definida — una que presenta como nuestra prioridad moral principal (al menos desde el punto de vista impersonal y secular) la reducción del riesgo existencial y el logro de una trayectoria civilizatoria que conduzca a un uso compasivo y jubiloso de los recursos cósmicos de la humanidad.

NOTAS

PREFACIO

1. No todas las notas finales contienen información útil, en cualquier caso.
2. No sé cuáles.

CAPÍTULO 1: DESARROLLOS DEL PASADO Y POSIBILIDADES DEL PRESENTE

1. A día de hoy la renta de subsistencia es de aproximadamente $ 400 (Chen y Ravallion, 2010). Un millón de ingresos de subsistencia son, pues, $ 400, 000,000. El producto bruto mundial actual es de aproximadamente $ 60.000.000.000.000 millones y en los últimos años ha crecido a un ritmo anual de alrededor del 4% (la tasa de crecimiento anual compuesta a partir de 1950, en base a Maddison [2010]). Estas cifras arrojan la estimación mencionada en el texto, que por supuesto es sólo una aproximación de cierto orden de magnitud. Si miramos directamente a las cifras de población, nos encontramos con que actualmente la población mundial necesita aproximadamente una semana y media para crecer en un millón; pero esto subestima el ritmo de crecimiento de la economía ya que el ingreso per cápita también estaría aumentando. En el 5000 a.C., tras la revolución agrícola, la población mundial estaba creciendo a un ritmo de alrededor de 1 millón por cada 200 años —una gran aceleración si lo comparamos con el ritmo de quizá un millón por cada millón años de principios de prehistoria humanoide— por lo que ya había tenido lugar entonces una gran aceleración. Aun así, es impresionante que la cantidad de crecimiento económico que hace siete mil años necesitaba 200 años, ahora sólo necesita noventa minutos, y el crecimiento de población mundial que entonces llevó dos siglos, ahora sólo requiere semana y media. Véase también Maddison (2005).
2. Este crecimiento y aceleración espectacular podría sugerir la idea de una posible llegada de la "singularidad", tal como esbozó John von Neumann en una conversación con el matemático Stanislaw Ulam:

   > Nuestra conversación se centró en los avances cada vez más acelerados de la tecnología y los cambios en el modo de vida humana, lo que hace parecer que nos acercamos a alguna singularidad esencial en la historia de la raza más allá de la cual los asuntos humanos, tal como los conocemos, no podrán continuar (Ulam, 1958).

3. Hanson (2000).
4. Vinge (1993); Kurzweil (2005).
5. Sandberg (2010).
6. Van Zanden (2003); Maddison (1999, 2001); De Long (1998).
7. Dos declaraciones optimistas habitualmente repetidas en la década de 1960: "Las máquinas serán capaces, dentro de veinte años, de hacer cualquier trabajo que un hombre pueda hacer" (Simon, 1965, 96); "Dentro de una generación... el problema de la creación de la inteligencia artificial estará sustancialmente resuelto"(Minsky, 1967, 2). Para una revisión sistemática de las predicciones sobre la IA, véase

Armstrong y Sotala (2012).

8. Véase, por ejemplo, Baum et al. (2011) y Armstrong y Sotala (2012).
9. Podría sugerirse, sin embargo, que los investigadores de IA saben menos acerca de los plazos de desarrollo de lo que ellos piensan —pero esto podría ir en ambos sentidos: podrían sobrestimar así como subestimar el momento en que se desarrollará la IA.
10. Good (1965, 33).
11. Una excepción es Norbert Wiener, quien sí tuvo reparos sobre las posibles consecuencias. Él escribió en 1960: "Si pudiéramos usar, para lograr nuestros propósitos, una agencia mecánica cuyo funcionamiento no pudiéramos interferir de manera eficiente una vez hubiéramos empezado, porque la acción fuera tan rápida e irrevocable que no tendríamos los datos para intervenir antes de que el la acción se hubiera completado, entonces sería mejor que nos aseguráramos de que el propósito dado a la máquina es el propósito que realmente deseamos y no simplemente una imitación colorida del mismo" (Wiener, 1960). Ed. Fredkin habló de sus preocupaciones sobre la IA superinteligente en una entrevista descrita en McCorduck (1979). En 1970, el mismo Good escribe acerca de los riesgos, e incluso pide la creación de una asociación para hacer frente a los peligros (Good [1970]; véase también el artículo posterior [Good, 1982] donde plantea algunas de las ideas de "normatividad indirecta "que se discuten en el Capítulo 13). En 1984, Marvin Minsky también estaba escribiendo acerca de muchas de estas preocupaciones decisivas (Minsky, 1984).
12. Cf. Yudkowsky (2008a). Sobre la importancia de evaluar las implicaciones éticas de las tecnologías del futuro potencialmente peligrosas antes de que sean factibles, véase Roache (2008).
13. McCorduck (1979).
14. Newell et al. (1959).
15. El programa SAINTS, el programa ANALOGY, y el programa STUDENT, respectivamente. Ver Slagle (1963), Evans (1964, 1968), y Bobrow (1968).
16. Nilsson (1984).
17. Weizenbaum (1966).
18. Winograd (1972).
19. Cope (1996); Weizenbaum (1976); Moravec (1980); Thrun et al. (2006); Buehler et al. (2009); Koza et al. (2003). El Departamento de Vehículos Motorizados de Nevada emitió la primera licencia para un coche sin conductor en mayo de 2012.
20. El sistema STANDUP (Ritchie et al., 2007).
21. Schwartz (1987). Schwartz está aquí caracterizando una visión escéptica que él pensaba que estaba presente en los escritos de Hubert Dreyfus.
22. Una voz crítica durante este período fue Hubert Dreyfus. Otros escépticos prominentes de esta época incluyen a John Lucas, Roger Penrose, y John Searle. Sin embargo, entre estos sólo Dreyfus estaba preocupado principalmente por refutar las afirmaciones sobre los logros prácticos que debíamos esperar de los paradigmas existentes de IA (a pesar de que parece haberse mantenido abierto a la posibilidad de que nuevos paradigmas podrían ir más lejos). El objetivo de Searle eran las teorías funcionalistas de filosofía de la mente, no los poderes instrumentales de sistemas de inteligencia artificial. Lucas y Penrose negaron que un ordenador clásico pudiera jamás llegar a ser programado para hacer todo lo que un matemático humano puede hacer, pero no negaron que cualquier función particular podría en principio ser automatizada o que las IAs podrían finalmente llegar a ser instrumentalmente muy poderosas. Cicerón remarcó que "no hay nada tan absurdo como para que no haya sido dicho ya por un filósofo" (Cicerón, 1923, 119); sin embargo, es sorprendentemente difícil pensar en cualquier pensador importante que haya negado la posibilidad de una superinteligencia artificial en el sentido utilizado en este libro.
23. En muchos sentidos, sin embargo, el aprendizaje que tiene lugar en una red neuronal es muy poco diferente del aprendizaje que tiene lugar en una regresión lineal, una técnica estadística desarrollada por Adrien-Marie Legendre y Carl Friedrich Gauss a principios de 1800.
24. El algoritmo básico fue descrito por Arthur Bryson y Yu-Chi Ho como un método de

optimización dinámica de varias etapas en el año 1969 (Bryson y Ho, 1969). La aplicación a redes neuronales fue sugerida por Paul Werbos en 1974 (Werbos, 1994), pero no fue hasta después de la obra de David Ru- melhart, Geoffrey Hinton y Ronald Williams de 1986 (Rumelhart et al., 1986) que el método comenzó a filtrarse poco a poco en la conciencia de una comunidad más amplia.

25. Las redes sin capas ocultas ya habían mostrado previamente tener una funcionalidad muy limitada (Minsky y Papert, 1969).
26. Por ejemplo, MacKay (2003).
27. Murphy (2012).
28. Pearl (2009).
29. Suprimimos varios detalles técnicos aquí a fin de no cargar innecesariamente la exposición. Tendremos ocasión de volver a examinar algunas de las cuestiones ignoradas en el Capítulo 12.
30. Un programa *p* es una descripción de la cadena x si *p,* ejecutada en alguna (particular) máquina universal de Turing U, exhibe v; escribimos esto como $U(P) =$ v. (La cadena v aquí representa un mundo posible). La complejidad de Kolmogorov de v es entonces $K(v): = \min_p\{l\ (p): U(p) =$ v$\}$, donde l *(p)* es la longitud de *p* en bits. La probabilidad de "Solomonoff" de v se define entonces como $M\ (v): =$

quier tipo de filosofía, escribir una gran novela de detectives, planear un golpe de Estado, o diseñar un importante nuevo producto de consumo.

61. Shapiro (1992).
62. Uno podría especular que una de las razones por las que ha sido difícil igualar las capacidades humanas en percepción, control motor, sentido común y comprensión del lenguaje es que nuestros cerebros tienen *wetware* dedicado a estas funciones —las estructuras neuronales se han optimizado durante grandes períodos de tiempo evolutivo. Por el contrario, el pensamiento lógico y las habilidades como jugar al ajedrez no son naturales para nosotros; así que quizás nos vemos obligados a recurrir a un grupo limitado de recursos cognitivos de propósito general para llevar a cabo estas tareas. Tal vez lo que nuestros cerebros hacen cuando nos involucramos en el razonamiento lógico explícito o en cálculo es de alguna manera análogo a la ejecución de una "máquina virtual", una simulación mental lenta y engorrosa de un ordenador de propósito general. Se podría decir entonces (un poco caprichosamente) que un programa de IA clásico no emula el pensamiento humano sino a la inversa: que un ser humano que piensa lógicamente está emulando un programa de IA.
63. Este ejemplo es controvertido: un punto de vista minoritario, representado por aproximadamente el 20% de los adultos en los EE.UU. y un número similar en muchos otros países desarrollados, sostiene que el Sol gira alrededor de la Tierra (Crabtree, 1999; Dean, 2005).
64. World Robotics (2011).
65. Estimado a partir de datos en Guizzo (2010).
66. Holley (2009).
67. También se usan enfoques estadísticos híbridos basados en reglas, pero actualmente son una pequeña parte del panorama.
68. Cross y Walker (1994); Hedberg (2002).
69. Basado en las estadísticas provenientes del Grupo TABB, una empresa de investigación en mercados de capital radicada en Nueva York-Londres (comunicación personal).
70. CFTC y SEC (2010). Para una perspectiva diferente sobre los acontecimientos del 6 de mayo de 2010, véase Grupo CME (2010).
71. Nada en el texto debe ser interpretado como un argumento en contra del intercambio algorítmico de alta frecuencia, que normalmente podría realizar una función beneficiosa al aumentar la liquidez y la eficiencia del mercado.
72. Un susto a un mercado más pequeño ocurrió el 1 de agosto de 2012, en parte debido a que el "interruptor" tampoco fue programado para detener el comercio en caso de acontecer cambios extremos en el número de acciones negociadas (Popper, 2012). Esto de nuevo adelanta otro tema que abordaremos más adelante: la dificultad de anticipar todas las formas específicas en que alguna regla de apariencia plausible podría pervertirse.
73. Nilsson (2009, 319).
74. Minsky (2006); McCarthy (2007); Beal y Winston (2009).
75. Peter Norvig, comunicación personal. Las clases de aprendizaje mediante máquinas son también muy populares, lo que refleja un bombo ortogonal respecto del "big data" (inspirado en, por ejemplo, Google y el Premio Netflix).
76. Armstrong y Sotala (2012).
77. Müller y Bostrom (de próxima publicación).
78. Véase Baum et al. (2011), otra encuesta citada allí y Sandberg y Bostrom (2011).
79. Nilsson (2009).
80. Esto está de nuevo condicionado a que no ocurra ninguna catástrofe destructora de la civilización. La definición de inteligencia artificial de nivel humano utilizada por Nilsson es "una IA capaz de realizar en torno al 80% de los trabajos tan bien o mejor que los humanos" (Kruel, 2012).
81. La tabla muestra los resultados de cuatro encuestas diferentes, así como los resultados combinados. Las dos primeras fueron encuestas realizadas en conferencias académicas: PT-AI, los participantes de la conferencia de *Filosofía y Teoría de la IA* en Salónica, 2011 (los encuestados fueron preguntados en noviembre de 2012), con una tasa de respuesta de 43 de 88; y AGI, los participantes de las conferencias de *Inteligencia Artificial General* y de los *Impactos y Riesgos de Inteligencia Artificial General,* ambas en Oxford, diciembre de 2012 (tasa de respuesta: 72/111). La encuesta EETN realizó un muestreo a los miembros

de la *Asociación Griega para la Inteligencia Artificial*, una organización profesional de conocidos investigadores sobre este campo, en abril de 2013 (tasa de respuesta: 26/250). La encuesta TOP100 obtuvo opiniones de los 100 mejores autores en inteligencia artificial medido por un índice de citas, en mayo 2013 (tasa de respuesta: 29/100).

82. Las entrevistas con 28 (en el momento de escribir) practicantes de IA y expertos relacionados han sido publicadas por Kruel (2011).
83. El diagrama muestra estimaciones normalizadas conforme a la media. Los medios son significativamente diferentes. Por ejemplo, las estimaciones medias para el resultado "extremadamente malo" fueron 7,6% (para el TOP100) y 17,2% (para el grupo combinado de asesores expertos).
84. Existe una abundante literatura que documenta la falta de fiabilidad de las previsiones de los expertos en muchos campos, y hay muchas razones para pensar que muchas de las conclusiones de estas investigaciones se aplican también al campo de la inteligencia artificial. En particular, los meteorólogos suelen tener demasiada confianza en sus predicciones, creyendo ser más exactos de lo que realmente son, y, por tanto, asignan muy poca probabilidad a la posibilidad de que su hipótesis preferida sea errónea (Tetlock, 2005). (Varios otros sesgos también se han documentado; véase, por ejemplo, Gilovich et al. [2002]). Sin embargo, la incertidumbre es un hecho ineludible de la condición humana, y muchas de nuestras acciones se basan inevitablemente en expectativas cuyas consecuencias a largo plazo son más o menos plausibles: en otras palabras, en predicciones probabilísticas. Negarse a ofrecer predicciones probabilísticas explícitas no haría desaparecer el problema epistémico; simplemente lo ocultaría a la vista (Bostrom, 2007). En cambio, deberíamos responder a la evidencia de cierto exceso de confianza ampliando nuestros intervalos de confianza (o "intervalos de credibilidad") —es decir, desmitificando nuestras funciones credenciales— y, en general, luchando lo mejor que podamos contra nuestros prejuicios, considerando diferentes perspectivas y buscando siempre la honestidad intelectual. A más largo plazo, también podemos trabajar para desarrollar técnicas, métodos de entrenamiento e instituciones que pudieran ayudar a lograr una mejor calibración. Ver también Armstrong y Sotala (2012).

CAPÍTULO 2: CAMINOS HACIA LA SUPERINTELIGENCIA

1. Esto se asemeja a la definición en Bostrom (2003c) y Bostrom (2006a). También puede compararse con la definición de Shane Legg ("La inteligencia mide la capacidad de un agente para lograr objetivos en una amplia gama de entornos") y sus formalizaciones (Legg, 2008). También es muy similar a la definición de Good de la ultrainteligencia en el capítulo 1 ("una máquina que pueda superar con creces todas las actividades intelectuales de cualquier hombre, sin importar lo inteligente que sea").
2. Por la misma razón, no se ofrece ninguna hipótesis acerca de si una máquina superinteligente podría tener "verdadera intencionalidad" (para Searle, podría; pero esto parece irrelevante para las preocupaciones de este libro). Y no tomamos ninguna posición en el debate internalismo / externalismo sobre el contenido mental que se ha estado librando en la literatura filosófica, o en la relacionada tesis de la mente extendida (Clark y Chalmers, 1998).
3. Turing (1950, 456).
4. Turing (1950, 456).
5. Chalmers (2010); Moravec (1976, 1988, 1998, 1999).
6. Véase Moravec (1976). Un argumento similar es adelantado por David Chalmers (2010).
7. Véase también Shulman y Bostrom (2012), donde estos temas están elaborados en más detalle.
8. Legg (2008) ofrece esta razón en apoyo a la afirmación de que el ser humano será capaz de recapitular el progreso de la evolución en períodos de tiempo mucho más cortos y con menores recursos computa- cionales (mientras señala que los recursos computacionales no ajustados de la evolución están lejos de su alcance). Baum (2004) sostiene que algunos

acontecimientos relacionados con la IA se produjeron antes, con la organización del genoma suponiendo una representación valiosa de los algoritmos evolutivos.

9. Whitman et al. (1998); Sabrosky (1952).
10. Schultz (2000).
11. Menzel y Giurfa (2001, 62); Truman et al. (1993).
12. Sandberg y Bostrom (2008).
13. Véase Legg (2008) para continuar el debate sobre este punto y sobre la promesa de que las funciones o entornos determinen la aptitud basada en un plácido panorama de tests de inteligencia pura.
14. Véase Bostrom y Sandberg (2009b) para una taxonomía y un análisis más detallado de las formas en que los ingenieros pueden superar la selección evolutiva histórica.
15. El análisis ha abordado el sistema nervioso de los seres vivos, sin referencia al coste de simular órganos o el entorno virtual que les rodea como parte de una función de aptitud. Es plausible que una función de aptitud adecuada podría poner a prueba la competencia de un organismo en particular en muchas menos operaciones de lo que se necesitaría para simular toda la computación neuronal del cerebro de ese organismo a lo largo de su ciclo de vida natural. Los programas de IA de hoy a menudo desarrollan y operan en ambientes muy abstractos (demostradores de teoremas en mundos simbólicos matemáticos, agentes en torneos mundiales de juegos simples, etc.).

    Un escéptico podría insistir en que un entorno abstracto sería insuficiente para la evolución de la inteligencia general, creyendo en cambio que el entorno virtual debería parecerse mucho al ambiente biológico real en el que nuestros ancestros evolucionaron. La creación de un mundo virtual físicamente realista requeriría mucha mayor inversión de recursos computacionales de simulación que un simple mundo de juguete o que un ámbito de problemas abstractos (en los que la evolución tendría acceso a un mundo real físico realista "gratuitamente"). En el caso límite, si se insistiera en la precisión micro-física completa, los requisitos computacionales podrían dispararse en proporciones ridículas. Sin embargo, ese pesimismo extremo casi seguro está injustificado; parece poco probable que el mejor entorno para la evolución de la inteligencia sea uno que imite a la naturaleza lo máximo posible. Es, por el contrario, plausible que fuera más eficiente utilizar un entorno de selección artificial, uno muy diferente al de nuestros antepasados, un entorno diseñado específicamente para promover adaptaciones que aumenten el tipo de inteligencia que buscamos evolucionar (razonamiento abstracto y habilidades para resolver problemas generales, por ejemplo, en contraposición a las reacciones instintivas máximamente rápidas o a un sistema visual altamente optimizado).

16. Wikipedia (2012b).
17. Para un tratamiento general de la teoría de la selección observacional, ver Bostrom (2002a). Para la aplicación específica a la cuestión presente, véase Shulman y Bostrom (2012). Para una corta introducción popular, véase Bostrom (2008b).
18. Sutton y Barto (1998, 21f); Schultz et al. (1997).
19. Este término fue introducido por Eliezer Yudkowsky; véase, por ejemplo, Yudkowsky (2007).
20. Éste es el escenario descrito por Good (1965) y Yudkowsky (2007). Sin embargo, también se podría considerar una alternativa en la que la secuencia iterativa tuviera algunas medidas que no implicaran un aumento de inteligencia, sino que simplificaran el diseño. Es decir, que en algunas etapas, la IA seminal pudiera reescribirse a sí misma con el fin de hacer que mejoras subsecuentes fueran más fácil de encontrar.
21. Helmstaedter et al. (2011).
22. Andres et al. (2012).
23. Adecuado para permitir formas instrumentalmente útiles de funcionamiento cognitivo y comunicación; pero todavía radicalmente empobrecido en comparación a la interfaz proporcionada por los músculos y órganos sensoriales de un cuerpo humano normal.
24. Sandberg (2013).

25. Véase la sección “Requisitos del sistema”, de Sandberg y Bostrom (2008, 79-81).
26. Un éxito de menor nivel podría ser una simulación del cerebro que tuviera una micro-dinámica biológicamente sugerente y que mostrara un amplio margen de actividad emergente típica de la especie, como un estado de sueño profundo o de actividad dependiente de la plasticidad. Aunque una simulación de este tipo pudiera ser un banco de pruebas útil para la investigación neurocientífica (mas uno podría estar cerca de plantear cuestiones éticas graves), no contaría como una emulación de cerebro completo a menos que la simulación fuera suficientemente precisa como para poder realizar una fracción sustancial del trabajo intelectual que era capaz de realizar el cerebro simulado. Como regla general, podríamos decir que para que una simulación de cerebro humano contara como una emulación de cerebro completo, tendría que ser capaz de expresar pensamientos verbales coherentes o tendría que tener la capacidad de aprender a hacerlo.
27. Sandberg y Bostrom (2008).
28. Sandberg y Bostrom (2008). Puede encontrarse una explicación más detallada en el informe original.
29. El primer mapa se ha descrito en Albertson y Thomson (1976) y en White et al. (1986). La red combinada (y en algunos casos corregida) está disponible en la web “WormAtlas” (http: // www.wormatlas.org/).
30. Para una revisión de los anteriores intentos de emular *c. elegans* y sus resultados, véase Kaufman (2011). Kaufman cita a un estudiante de doctorado ambicioso que trabajaba en el área, David Dal- rymple, diciendo: “Con las técnicas de optogenética, estamos justo en un punto en el que no es una propuesta escandalosa el alcanzar la capacidad de leer y escribir en cualquier sistema nervioso de una *c. elegans* viva, usando un sistema automatizado de alto rendimiento... Espero terminar con la *c. elegans* en 2-3 años. Yo estaría muy sorprendido, valga mi predicción lo que valga, que esto siguiera siendo un problema abierto en el año 2020 “(Dalrymple, 2011). Los modelos cerebrales en busca de realismo biológico que estaban codificados a mano (no generados automáticamente) han logrado algunas funciones básicas; véase, por ejemplo, Eliasmith et al. (2012).
31. La *caenorhabditis elegans* tiene algunas propiedades especiales convenientes. Por ejemplo, el organismo es transparente, y el patrón de cableado de su sistema nervioso no cambia de un individuo a otro.
32. Si el resultado final es la IA neuromórfica en lugar de la emulación de cerebro completo, entonces podría ser o no ser el caso que los conocimientos pertinentes se deriven de intentos de simular el cerebro humano. Es concebible que los trucos corticales importantes sean descubiertos durante el estudio de cerebros animales (no humanos). Trabajar con algunos cerebros animales podría ser más fácil que con cerebros humanos, y los cerebros más pequeños requerirían menos recursos para escanearlos y modelarlos. La investigación sobre el cerebro de los animales también estaría sujeta a menos regulación. Incluso es posible que la primera inteligencia artificial de nivel humano se creara completando una emulación de cerebro completo de un animal adecuado y encontrando entonces formas de mejorar las mentes digitales resultantes. Por lo tanto la humanidad podría recibir su merecido a manos de una rata de laboratorio o de un macaco supermejorado.
33. Uauy y Dangour (2006); Georgieff (2007); Stewart et al. (2008); Eppig et al. (2010); Cotman y Berchtold (2002).
34. Según la Organización Mundial de la Salud, en 2007 cerca de dos mil millones de individuos tienen una ingesta insuficiente de yodo (The Lancet, 2008). La deficiencia grave de yodo obstaculiza el desarrollo neurológico y conduce al cretinismo, que implica una pérdida promedio de alrededor de 12,5 puntos de coeficiente intelectual (Qian et al., 2005). La condición se puede prevenir fácilmente y de manera económica mediante un enriquecimiento en sal de la dieta (Horton et al., 2008).
35. Bostrom y Sandberg (2009a).
36. Bostrom y Sandberg (2009b). Un aumento típico del rendimiento esperado proveniente de mejoras farmacológicas y nutricionales se encuentra en el rango de 10-20% en tests que miden la memoria la atención, etc. Pero en general es dudoso que tales ganancias sean reales, sostenibles a más largo plazo, e indicativas de resultados correspondientemente mejorados en situaciones problemáticas del mundo real (Repantis et al., 2010). Por

ejemplo, en algunos casos puede haber un deterioro de compensación en algunas dimensiones del rendimiento que no son medidas por los tests (Sandberg y Bostrom, 2006).

37. Si hubiera una manera fácil de mejorar la cognición, uno esperaría que la evolución ya se hubiera aprovechado de ella. En consecuencia, el tipo más prometedor de nootrópico para investigar puede ser uno que prometa aumentar la inteligencia en maneras que habrían reducido nuestra aptitud en un entorno para ancestral —por ejemplo, aumentando el tamaño de la cabeza al nacer o amplificando el metabolismo de la glucosa en el cerebro. Para una discusión más detallada sobre esta idea (junto con varias especificaciones importantes), véase Bostrom (2009b).
38. Los espermatozoides son más difíciles de mostrar en pantalla porque, a diferencia de los embriones, consisten en sólo una célula y —una célula debe ser destruida para hacer la secuenciación. Los ovocitos también consisten en una sola célula; sin embargo, la primera y segunda división celular son asimétricas y producen células hijas con muy poco citoplasma, el cuerpo polar. Ya que los cuerpos polares contienen el mismo genoma de la célula principal y son redundantes (finalmente degeneran) se les puede hacer una biopsia y ser utilizados para ser analizados en pantalla (Gianaroli, 2000).
39. Cada una de estas prácticas fue objeto de cierta controversia ética cuando se presentaron, pero parece que hay una tendencia hacia una creciente aceptación. Las actitudes hacia la ingeniería genética humana y la selección de embriones varían significativamente en todas las culturas, lo que sugiere que el desarrollo y aplicación de nuevas técnicas probablemente se llevará a cabo incluso si algunos países adoptan inicialmente una postura cautelosa, aunque la velocidad a la que esto sucede estará influenciado por presiones morales, religiosas, y políticas.
40. Davies et al. (2011); Benyamin et al. (2013); Plomin et al. (2013). Véase también Mardis (2011); Hsu (2012).
41. La heredabilidad en un amplio sentido del CI adulto está habitualmente estimada en un rango de 0,50,8 dentro de los estratos de clase media de los países desarrollados (Bouchard, 2004, 148). La heredabi- lidad en sentido estrecho, que mide la parte de la varianza que es atribuible a factores genéticos aditivos, es inferior (en el rango de 0,3-0,5), pero sigue siendo considerable (Devlin et al., 1997; Davies et al., 2011; Visscher et al., 2008). Estas estimaciones podrían cambiar para diferentes poblaciones y entornos, pues ya se está estudiando cómo las heredabilidades varían en función de la población y el medio ambiente. Por ejemplo, las heredabilidades más bajas se han encontrado entre los niños procedentes de entornos desfavorecidos (Benyamin et al., 2013; Turkheimer et al., 2003). Nisbett et al. (2012) revisa numerosas influencias ambientales sobre la variación de la capacidad cognitiva.
42. Los siguientes párrafos se basan en gran medida en el trabajo conjunto con Carl Shulman (Shulman y Bostrom, 2014).
43. Esta tabla está tomada de Shulman y Bostrom (2014). Se basa en un modelo de juguete que asume una distribución gaussiana de CIs anticipados entre embriones con una desviación estándar de 7,5 puntos. La cantidad de mejora de la cognición que se reparta en los diferentes embriones depende de cómo los embriones se diferencian unos de los otros en las variantes genéticas aditivas cuyos efectos conocemos. Los hermanos tienen un coeficiente de relación de V2, y variantes genéticas aditivas comunes representan la mitad o menos de la varianza en la inteligencia fluida adulta (Davies et al., 2011). Estos dos hechos sugieren que, cuando la desviación estándar poblacional observada en los países desarrollados es de 15 puntos, la desviación estándar de las influencias genéticas dentro de un lote de embriones sería de 7,5 puntos o menos.
44. Con información imperfecta acerca de los efectos genéticos aditivos sobre la capacidad cognitiva, se reducirían los tamaños del efecto. Sin embargo, incluso una pequeña cantidad de conocimiento ayudaría mucho, porque las ganancias derivadas de la selección no escalan de manera lineal con la porción de variación que podemos predecir. En su lugar, la eficacia de nuestra selección depende de la predicción en la desviación estándar de CI medio, que escala como la raíz cuadrada de la variación. Por ejemplo, si se pudiera dar cuenta del 12,5% de la variación, esto podría ofrecer efectos medios tan grandes como los de la Tabla 1, que asumen el 50%. Para comparar, un estudio reciente (Rietveld et al., 2013) afirma que ya tienen identificado un 2,5% de variación.
45. En comparación, una práctica habitual hoy en día consiste en la creación de menos de

diez embriones.

46. Las células madre adultas y embrionarias pueden ser inducidas a convertirse en células de esperma y ovocitos, que luego se pueden fusionar para producir un embrión (Nagy et al., 2008; Nagy y Chang, 2007). Los óvulos precursores también pueden formar blastocistos partenogenéticos, embriones fertilizados y no viables, capaces de producir líneas de células madre embrionarias para el proceso (Mai et al., 2007).
47. La opinión es de Katsuhiko Hayashi, según se informa en Cyranoski (2013). El Grupo Hinxton, un consorcio internacional de científicos que analiza la ética y los desafíos de las células madre, predijo en 2008 que los gametos derivados de células madre humanas estarían disponibles dentro de diez años (Grupo Hinxton, 2008), y los desarrollos hasta ahora son ampliamente consistentes con esto.
48. Sparrow (2013); Miller (2012); The Uncertain Future (2012).
49. Sparrow (2013).
50. Las preocupaciones seculares podrían centrarse en los impactos previstos en desigualdad social, la seguridad médica del procedimiento, los temores de una "carrera de ratas" mejorativa, los derechos y responsabilidades de los padres respecto de su descendencia, la sombra de la eugenesia del siglo XX, el concepto de dignidad humana y los límites propios de la participación de los Estados en las decisiones reproductivas de sus ciudadanos. (Para una discusión sobre la ética de la mejora cognitiva, ver Bostrom y Ord [2006], Bostrom y Roache [2011], y Sandberg y Savulescu [2011]). Algunas tradiciones religiosas pueden ofrecer preocupaciones adicionales, incluidas las que se centran en el estatus moral de los embriones o en los límites propios de la acción humana dentro del esquema de la creación.
51. Para evitar los efectos negativos de la endogamia, la selección de embriones por iteración requeriría o bien un gran suministro de partida por parte de los donantes o el gasto de sustancial energía selectiva para reducir alelos recesivos perjudiciales. Cualquier alternativa tendería a empujar a que la descendencia tuviera una relación genética menos estrecha con sus padres (y estuvieran más relacionados entre sí).
52. Adaptado de Shulman y Bostrom (2014).
53. Bostrom (2008b).
54. Aún no se sabe cómo de difícil serán los obstáculos epigenéticos (Chason et al., 2011; Iliadou et al., 2011).
55. Mientras que la capacidad cognitiva es un rasgo bastante heredable, puede que haya pocos o ningún alelo común o polimórfico que tenga individualmente un gran efecto positivo sobre la inteligencia (Da- vis et al., 2010; Davies et al., 2011; Rietveld et al., 2013). A medida que los métodos de secuenciación mejoren, la asignación de alelos de baja frecuencia y sus correlatos cognitivos y de comportamiento serán cada vez más factibles. Hay algunas pruebas teóricas que sugieren que algunos alelos que causan trastornos genéticos en los homocigotos pueden proporcionar ventajas cognitivas considerables a los portadores de heterocigotos, lo que lleva a la predicción de que los heterocigotos de pacientes de Gaucher, Tay-Sachs, y Niemann-Pick tendrían unos 5 puntos más de CI respecto de los grupos de control (Cochran et al., 2006). El tiempo dirá si esto es así.
56. Un artículo (Najman y Crowell, 2000) estima 175 mutaciones por genoma en cada generación. Otro (Lynch, 2010), utilizando diferentes métodos, estima que el recién nacido promedio tiene entre 50 y 100 nuevas mutaciones, y Kong et al. (2012) supone una cifra de alrededor de 77 nuevas mutaciones por generación. La mayoría de estas mutaciones no afectan al funcionamiento, o lo hacen sólo en un grado imperceptiblemente pequeño; pero los efectos combinados de muchas mutaciones perjudiciales muy pequeñas podrían implicar una importante pérdida de aptitud. Véase también Crow (2000).
57. Crow (2000); Lynch (2010).
58. Hay algunas advertencias potencialmente importantes sobre esta idea. Es posible que el genoma modal necesite algunos ajustes con el fin de evitar problemas. Por ejemplo, partes del genoma pueden ser adaptadas a la interacción con otras partes bajo el supuesto de que todas las piezas funcionen con un cierto nivel de eficiencia. Aumentar la eficiencia de las partes podría entonces conducir a un exceso a lo largo de algunas rutas metabólicas.
59. Estos compuestos fueron creados por Mike Mike a partir de fotografías individuales

tomadas por Virtual Flavius (Mike, 2013).

60. Éstas podrían, por supuesto, tener algunos efectos más rápidos —por ejemplo, cambiando las expectativas de la gente sobre lo que está por venir.
61. Louis Harris & Associates (1969); Mason (2003).
62. Kalfoglou et al. (2004).
63. Los datos son, obviamente, limitados, pero los individuos seleccionados para los resultados 1 entre 10.000 en las pruebas de habilidad de la infancia han demostrado, en estudios longitudinales, ser mucho más propensos a convertirse en profesores titulares, conseguir patentes y tener más éxito en los negocios que los que tienen puntuaciones ligeramente menos excepcionales (Kell et al., 2013). Roe (1953) estudió a sesenta y cuatro eminentes científicos y encontró una capacidad cognitiva desviada en un promedio de tres a cuatro por encima de la norma de la población y sorprendentemente más alta de lo que es típico entre los científicos en general. (La capacidad cognitiva también está correlacionada con las ganancias de una vida y con resultados no financieros tales como la esperanza de vida, las tasas de divorcio y la probabilidad de abandonar la escuela [Deary, 2012]). Un desplazamiento hacia arriba de la distribución de la capacidad cognitiva tendría desproporcionadamente grandes efectos en las partes inferiores, sobre todo aumentando el número de superdotados y reduciendo el número de personas con retraso y dificultades de aprendizaje. Véase también Bostrom y Ord (2006) y Sandberg y Savulescu (2011).
64. Por ejemplo, Warwick (2002). Stephen Hawking llegó a sugerir que dar este paso podría ser necesario con el fin de mantenerse al día con los avances en inteligencia artificial: "Tenemos que desarrollar lo más rápidamente posible las tecnologías que hacen posible una conexión directa entre el cerebro y el ordenador, para que los cerebros artificiales contribuyan a la inteligencia humana en lugar de oponerse a ella" (reportado en Walsh [2001]). Ray Kurzweil está de acuerdo: "En lo concerniente a la recomendación de Hawking... es decir, a la conexión directa entre el cerebro y los ordenadores, estoy de acuerdo en que esto es razonable, deseable e inevitable. [sic] Ha sido mi recomendación durante años" (Kurzweil, 2001).
65. Véase Lebedev y Nicolelis (2006); Birbaumer et al. (2008); Mak y Wolpaw (2009); y Nicolelis y Lebedev (2009). Una visión más personal sobre el problema de la mejora a través de implantes se puede encontrar en Chorost (2005, cap. 11).
66. Smeding et al. (2006).
67. Degnan et al. (2002).
68. Dagnelie (2012); Shannon (2012).
69. Perlmutter y Mink (2006); Lyons (2011).
70. Koch et al. (2006).
71. Schalk (2008). Para una revisión general del estado actual de la técnica, véase Berger et al. (2008). En referencia al tesis de que esto nos ayudaría a llegar a una inteligencia mejorada, véase Warwick (2002).
72. Algunos ejemplos: Bartels et al. (2008); Simeral et al. (2011); Krusienski y Shih (2011); y Pasqualotto et al. (2012).
73. Por ejemplo, Hinke et al. (1993).
74. Hay excepciones parciales a esto, especialmente en el procesamiento sensorial temprano. Por ejemplo, la corteza visual primaria utiliza un mapeo retinotópico, lo que más o menos significa que los conjuntos neuronales adyacentes reciben estímulos de las zonas adyacentes a las retinas (aunque las columnas de dominio ocular complican un tanto el mapeo).
75. Berger et al. (2012); Hampson et al. (2012).
76. Algunos implantes cerebrales requieren dos formas de aprendizaje: el dispositivo de aprendizaje destinado a interpretar las representaciones neurales del organismo y el aprendizaje orgánico para utilizar el sistema mediante la generación de patrones de disparo neuronal apropiados (Carmena et al., 2003).
77. Se ha sugerido que deberíamos considerar a las personas jurídicas (empresas, sindicatos, gobiernos, iglesias, etc.) como agentes inteligentes artificiales, entidades con sensores y efectores, capaces de representar conocimiento y realizar inferencias y tomar medidas (por ejemplo, Kuipers [2012]; cf. Huebner [2008] para una discusión sobre la posibilidad de que existan representaciones colectivas). Son claramente poderosas y ecológicamente exitosas, aunque sus capacidades y estados

internos son diferentes de los de los humanos.

78. Hanson (1995, 2000); Berg y Rietz (2003).
79. En el lugar de trabajo, por ejemplo, los empleadores podrían utilizar detectores de mentiras para acabar con el robo y el escaqueo de los empleados, preguntando al empleado al final de cada día si había robado algo y si había trabajado tan duro como hubiera podido. De igual manera, se les puede preguntar a los líderes políticos y empresariales si estaban buscando de todo corazón el interés de sus accionistas o participantes. Los dictadores podrían utilizarlos para localizar a generales sediciosos dentro del régimen o a los responsables de presuntos problemas entre la población en general.
80. Uno podría imaginar técnicas de neuroimagen que permitieran detectar firmas neurales motivadas por cognición. Sin detección de autoengaño, la detección de mentiras favorecería a individuos que creen sus propias mentiras. Mejores pruebas para los tests de auto-engaño también podrían ser utilizadas para entrenar la racionalidad y para el estudio de la efectividad de las intervenciones dirigidas a reducir el sesgo.
81. Bell y Gemmel (2009). Un primer ejemplo se encuentra en la obra de Deb Roy del MIT, quien registró cada momento de los tres primeros años de vida de su hijo. El análisis de estos datos audiovisuales están dando información sobre el desarrollo del lenguaje; ver Roy (2012).
82. El crecimiento en el total mundial de la población de seres humanos biológicos contribuirá sólo como un pequeño factor. Los escenarios que implican inteligencia artificial podrían presentar una explosión de población mundial (incluyendo mentes digitales) de muchos órdenes de magnitud en un breve período de tiempo. Pero ese camino hacia la superinteligencia presupone la inteligencia artificial o la emulación de cerebro completo, por lo que no es necesario tenerlo en cuenta en esta subsección.
83. Vinge (1993).

CAPÍTULO 3: FORMAS DE SUPERINTELIGENCIA

1. Vernor Vinge ha utilizado el término “superinteligencia débil” para referirse a este tipo de mentes humanas aceleradas (Vinge, 1993).
2. Por ejemplo, si un sistema muy rápido pudiera hacer todo lo que puede hacer cualquier humano excepto bailar una mazurca, todavía deberíamos llamarlo una superinteligencia de velocidad. Nuestro interés se centra en esas capacidades cognitivas básicas que tienen importancia económica o estratégica.
3. Al menos una aceleración de un millón de veces más que los cerebros humanos es físicamente posible, como se puede ver teniendo en cuenta la diferencia en la velocidad y energía de los procesos relevantes del cerebro en comparación con el procesamiento de información más eficiente. La velocidad de la luz es más de un millón de veces mayor que la de la transmisión neuronal, los enlaces sinápticos disipan un millón de veces más calor del que es termodinámicamente necesario, y las frecuencias de radio actuales son un millón de veces más rápidas que las frecuencias neuronales (Yudkowsky [2008a]; véase también Drexler [1992]). Los límites últimos de la superinteligencia de velocidad están limitados por retardos en la comunicación a la velocidad de la luz, por límites cuánticos sobre la velocidad de las transiciones de estado, y por el volumen necesario para contener la mente (Lloyd, 2000). El “portátil definitivo” descrito por Lloyd (2000) se ejecutaría a $1{,}4 \times 10^{21}$ FLOPS, una emulación de cerebro acelerada por $3{,}8 \times 10^{29}$ (asumiendo que la emulación pudiera estar lo suficientemente paralelizada). La construcción de Lloyd, sin embargo, no pretende ser tecnológicamente creíble; sólo pretende ilustrar que las restricciones en computación pueden derivarse fácilmente de las leyes físicas básicas.
4. Con las emulaciones, existe también una cuestión de cuánto tiempo puede una mente similar a la humana seguir trabajando en algo antes de volverse loca o caer en la rutina. Incluso con variedad de tareas y vacaciones

regulares, no es seguro que una mente similar a la humana pudiera vivir durante miles de años subjetivos sin desarrollar problemas psicológicos. Por otra parte, si la capacidad total de la memoria es limitada —una consecuencia de tener una población limitada de neuronas— entonces el aprendizaje acumulativo no puede continuar indefinidamente: más allá de cierto punto, la mente debe empezar a olvidar una cosa por cada cosa nueva que aprende. (La inteligencia artificial podría ser diseñada para paliar estos potenciales problemas).

5. En consecuencia, los nanomecanismos que se mueven a un modesto 1 m/s habitualmente tienen escalas temporales de nanosegundos. Véase la sección 2.3.2 de Drexler (1992). Robin Hanson menciona cuerpos de 7 mm del robot "tinkerbell" que se mueven 260 veces más rápido respecto de la velocidad normal (Hanson, 1994).
6. Hanson (2012).
7. La "inteligencia colectiva" no se refiere a la paralelización a pequeña escala de hardware de computación, sino a la paralelización a nivel de agentes autónomos inteligentes como los seres humanos. La im- plementación de una sola emulación en una máquina paralela masiva podría resultar en una superin- teligencia de velocidad si el ordenador paralelo fuera suficientemente rápido: pero no produciría una inteligencia colectiva.
8. Mejoras en la velocidad o en la calidad de los componentes individuales también podrían afectar indirectamente el funcionamiento de la inteligencia colectiva, pero aquí tenemos en cuenta principalmente mejoras englobadas bajo las otras dos formas de superinteligencia de nuestra clasificación.
9. Se ha argumentado que una densidad más alta de población desencadenó la revolución del paleolítico superior y que cuando se llegó a un cierto umbral de acumulación de complejidad cultural fue algo mucho más sencillo (Powell et al., 2009).
10. ¿Qué pasa con internet? No parece haber supuesto un impulso de gran tamaño. Quizás lo haga eventualmente. Tomó siglos o milenios para que los otros ejemplos enumerados aquí revelaran todo su potencial.
11. Esto, obviamente, no pretende ser un experimento mental realista. Un planeta lo suficientemente grande como para sostener siete mil billones de organismos humanos con la tecnología actual podría im- plosionar, a menos que estuviera hecho de una materia muy ligera o estuviera hueco y fuera sostenido por la presión o por otros medios artificiales. (Una esfera de Dyson o un planeta-caparazón podrían ser mejores soluciones). La historia se hubiera desarrollado de manera diferente en una superficie tan vasta. Dejemos todo esto de lado.
12. Nuestra atención se centra en las propiedades funcionales de una inteligencia unificada, no en la cuestión de si tal intelecto tendría cualidades o si sería una mente en el sentido de experimentar consciencia subjetiva. (Uno podría pensar, sin embargo, qué tipo de experiencia consciente podría derivarse de intelectos que estuvieran más o menos integrados que los cerebros humanos. Algunos puntos de vista sobre la conciencia, como la teoría del espacio de trabajo mundial, creen que podría esperarse que cerebros más integrados tuvieran una conciencia más amplia. Cf. Baars [1997], Shanahan [2010], y Schwitzgebel [2013]).
13. Incluso pequeños grupos de humanos que hubieran permanecido aislados durante algún tiempo todavía podrían beneficiarse de los resultados intelectuales de una inteligencia colectiva más grande. Por ejemplo, el lenguaje que utilizan podría haber sido desarrollado por una comunidad lingüística mucho más grande, y las herramientas que utilizan podrían haber sido inventadas en una población mucho mayor antes de que el pequeño grupo se aislara. Pero incluso si un pequeño grupo siempre hubiera estado aislado, aún podría ser parte de una inteligencia colectiva más grande de lo que parece a simple vista —es decir, la inteligencia colectiva incluye no sólo a la presente, sino a todas las generaciones ancestrales, un agregado que pueden funcionar como un sistema de

procesamiento de información volcado hacia adelante.

14. Gracias a la tesis de Church-Turing, todas las funciones computables son computables por una máquina de Turing. Ya que cualquiera de las tres formas de superinteligencia podría simular una máquina de Turing (si se les diera acceso a memoria ilimitada y se les dejara operar indefinidamente), en base a este criterio formal son computacionalmente equivalentes. De hecho, un ser humano promedio (siempre con papel ilimitado y sin límite de tiempo) también podría implementar una máquina de Turing, y, por lo tanto, también sería equivalente en base a ese criterio. Lo que importa para nuestros propósitos, sin embargo, es lo que estos diferentes sistemas pudieran lograr en la práctica, con memoria finita y en un plazo razonable. Y las variaciones de eficiencia son tan grandes que uno puede fácilmente hacer algunas distinciones. Por ejemplo, a un individuo típico con un CI de 85 se le puede enseñar a imple- mentar una máquina de Turing. (Posiblemente, aún podría ser posible entrenar a algunos chimpancés particularmente dotados y dóciles para que hicieran esto). Sin embargo, a todos los efectos prácticos, tal individuo es presumiblemente incapaz de, por ejemplo, desarrollar de forma independiente la teoría de la relatividad general o de ganar una medalla Fields.
15. Las tradiciones orales de narración pueden producir grandes obras (como las epopeyas homéricas), pero tal vez algunos de los autores contribuyentes poseían dones poco comunes.
16. A menos que contenga como componentes a intelectos que tuvieran inteligencia de velocidad o de calidad.
17. Nuestra incapacidad para especificar en qué consisten todos estos problemas puede deberse en parte a no haberlo intentado: no tiene mucho sentido pasarse el tiempo detallando trabajos intelectuales que ningún individuo ni organización actualmente factible puede realizar. Pero también es posible que incluso conceptualizar algunos de estos puestos de trabajo es en sí mismo uno de esos trabajos para los que actualmente carecemos de los cerebros necesarios.
18. Cf. Boswell (1917); véase también Walker (2002).
19. Esto ocurre principalmente en ráfagas cortas de subconjuntos de neuronas —la mayoría tienen tasas de disparo más tranquilas (Gray y McCormick, 1996; Steriade et al., 1998). Hay algunas neuronas ("neuronas castañeantes", también conocidas como células de "estallido rítmicamente rápido") que pueden alcanzar frecuencias de disparo de hasta 750 Hz, pero éstos parecen ser valores atípicamente extremos.
20. Feldman y Ballard (1982).
21. La velocidad de conducción depende del diámetro del axón (axones más gruesos son más rápidos) y de si el axón está mielinizado. Dentro del sistema nervioso central, los retardos de transmisión pueden variar desde menos de un milisegundo hasta 100 ms (Kandel et al., 2000). La transmisión en fibras ópticas es de alrededor de un 68% c (debido al índice de refracción del material). Los cables eléctricos tienen más o menos la misma velocidad, 59-77% c.
22. Esto supone una velocidad de señal de 70% c. Suponer un 100% c eleva la estimación a un $1{,}8 \times 10^{18}$ m^3.
23. El número de neuronas en un cerebro adulto masculino humano se ha estimado en $86{,}1 \pm 8{,}1$ miles de millones, un número al que se llegó disolviendo cerebros y fraccionando los núcleos celulares, contando las que estaban teñidas con un marcador específico de neuronas. En el pasado, las estimaciones en torno a los 75-125 miles de millones de neuronas eran comunes. Estos normalmente se basaban en contar manualmente la densidad de células en pequeñas regiones representativas (Azevedo et al., 2009).
24. Whitehead (2003).
25. Los sistemas de procesamiento de información pueden muy probablemente utilizar procesos a escala molecular para la computación y almacenamiento

de datos y llegar a un tamaño al menos planetario de extensión. Los límites físicos al cálculo establecidos por la mecánica cuántica, la relatividad general y la termodinámica están, sin embargo, mucho más allá de ese nivel de "cerebro-Júpiter" (Sandberg, 1999; Lloyd, 2000).

26. Stansberry y Kudritzki (2012). La electricidad utilizada en los centros de datos en todo el mundo asciende a 1,1-1,5% del consumo total de electricidad (Koomey, 2011). Veáse también Muehlhauser y Salamon (2012).
27. Esto es una simplificación excesiva. El número de pedazos que una memoria de trabajo puede mantener depende tanto de la información como de la tarea; sin embargo, está claramente limitada a un pequeño número de pedazos. Ver Miller (1956) y Cowan (2001).
28. Un ejemplo podría ser que la dificultad de aprender conceptos booleanos (categorías definidas por reglas lógicas) fuera lógicamente proporcional a la longitud de la fórmula proposicional equivalente más corta. Normalmente, incluso fórmulas de literalmente una longitud de sólo 3-4 de largo son muy difíciles de aprender. Ver Feldman (2000).
29. Ver Landauer (1986). Este estudio se basa en estimaciones experimentales de las tasas de aprendizaje y olvido en los seres humanos. Teniendo en cuenta que el aprendizaje implícito podría hacer variar la estimación un poco. Si se asume una capacidad de almacenamiento de $\sim$ 1 bit por sinapsis, se obtiene un límite superior de la capacidad memorística humana de alrededor de 10^{15} bits. Para una visión general de las diferentes estimaciones, véase el Apéndice A de Sandberg y Bostrom (2008).
30. El canal del ruido puede desencadenar potenciales de acción, y el ruido sináptico produce una variabilidad significativa en la intensidad de las señales transmitidas. Los sistemas nerviosos parecen haber evolucionado para hacer numerosas compensaciones entre la tolerancia al ruido y sus costes (masa, tamaño, demoras de tiempo); ver Faisal et al. (2008). Por ejemplo, los axones no pueden ser más delgados que 0,1 micras sin que la apertura al azar de los canales iónicos puedan desencadenar potenciales de acción espontáneos (Faisal et al., 2005).
31. Trachtenberg et al. (2002).
32. En términos de memoria y potencia de cálculo, aunque no en términos de eficiencia energética. El ordenador más rápido del mundo en el momento de la escritura es el "Tianhe-2" chino, que desplazó a Cray Inc. Titan en junio de 2013 con un rendimiento de 33.86 petaFLOPS. Utiliza 17,6 MW de potencia, casi seis órdenes de magnitud más que los $\sim$ 20 W. del cerebro
33. Téngase en cuenta que esta encuesta sobre fuentes de ventaja de las máquinas es *disyuntiva:* nuestro argumento tiene éxito incluso si algunos de los elementos que se enumeran son ilusorios, siempre y cuando haya al menos una fuente que pueda proporcionar una ventaja suficientemente grande.

CAPÍTULO 4: LA CINÉTICA DE UNA EXPLOSIÓN DE INTELIGENCIA

1. El sistema puede no llegar a una de estas líneas de base en ningún punto claramente definido. Puede en su lugar haber un intervalo durante el cual el sistema se vuelva gradualmente capaz de superar al equipo de investigación externo en un número creciente de tareas de desarrollo y mejoras del sistema.
2. En el último medio siglo, ha habido al menos un escenario que se ha creído podría llevar a acabar con el orden mundial existente en pocos minutos u horas: una guerra termonuclear global.
3. Esto sería coherente con la observación de que el efecto de Flynn —el aumento secular en las puntuaciones de CI en la mayoría de poblaciones a un ritmo de unos 3 puntos de CI por década en los últimos 60 años más o menos— parece haber cesado o incluso revertido en los últimos años en algunos países altamente desarrollados como el Reino Unido, Dinamarca y Noruega (Teasdale y Owen, 2008; Sundet et al., 2004). La causa del efecto Flynn en el pasado —y la cuestión de hasta qué punto representa algún

beneficio real en inteligencia general o simplemente una mejor habilidad en la resolución de tests de CI— ha sido objeto de amplio debate y aún no se sabe a ciencia cierta. Incluso si el efecto Flynn reflejara (al menos parcialmente) ganancias cognitivas reales, e incluso si el efecto estuviera disminuyendo o incluso revirtiendo, esto no prueba que aún hayamos alcanzado rendimientos decrecientes en la causa subyacente que fuera responsable del efecto Flynn en el pasado. La disminución o reversión podría deberse a algún factor perjudicial independiente que de otro modo habría producido un declive aún más grande.

4. Bostrom y Roache (2011).
5. La terapia génica somática podría eliminar el retraso madurativo, pero es técnicamente mucho más difícil que las intervenciones en la línea germinal y tiene un potencial máximo más pequeño.
6. El promedio de crecimiento de la productividad de la economía mundial por año durante el período 1960-2000 fue de 4,3% (Isaksson, 2007). Sólo una parte de este crecimiento de la productividad se debe al aumento de la eficiencia organizacional. Algunas redes o procesos organizativos están mejorando claramente a un ritmo mucho más rápido.
7. La evolución del cerebro biológico estuvo sujeta a muchas restricciones y compensaciones que se rebajarían drásticamente si la mente se trasladara a un medio digital. Por ejemplo, el tamaño del cerebro está limitado por el tamaño de la cabeza, y una cabeza que fuera demasiado grande tendría problemas para pasar a través del canal de parto. Un cerebro grande también absorbe recursos metabólicos y es un peso muerto que impide el movimiento. La conectividad entre ciertas regiones del cerebro podría estar limitada por restricciones estéricas —el volumen de materia blanca es significativamente mayor que el volumen de materia gris que se encarga de conectar. La disipación de calor está limitada por el flujo de sangre, y podría estar cerca del límite superior de un funcionamiento aceptable. Además, las neuronas biológicas son ruidosas, lentas, y necesitadas de una protección constante, el mantenimiento y el reabastecimiento a través de las células gliales y de los vasos sanguíneos (que contribuye a la aglomeración intracraneal). Véase Bostrom y Sandberg (2009b).
8. Yudkowsky (2008a, 326). Para un análisis más reciente, véase Yudkowsky (2013).
9. La imagen muestra la capacidad cognitiva como un parámetro de una sola dimensión, para no complicar el dibujo. Pero esto no es esencial para el punto que aquí se defiende. Uno podría, por ejemplo, representarse en su lugar un perfil de la capacidad cognitiva como una hipersuperficie en un espacio multidimensional.
10. Lin et al. (2012).
11. Se obtiene un cierto aumento de inteligencia colectiva simplemente aumentando el número de inteligencias constituyentes. Si se hace, por lo menos debería permitir un mejor rendimiento general en tareas que pudieran paralelizarse fácilmente. Para cosechar todos los rendimientos de una explosión de población, sin embargo, uno también tendría que lograr algún (más que mínimo) nivel de coordinación entre los constituyentes.
12. La distinción entre la inteligencia de velocidad y de calidad es, de todos modos, borrosa en el caso de sistemas de IA no-neuromórficos.
13. Rajab et al. (2006, 41-52).
14. Se ha sugerido que el uso de circuitos integrados configurables (FPGAs) en lugar de procesadores de propósito general podría aumentar las velocidades computacionales de simulaciones de redes neuronales en hasta dos órdenes de magnitud (Markram, 2006). Un estudio de alta resolución de la modelización del clima en un rango de petaFLOP encontró una reducción de costes de entre veinticuatro y treinta y cuatro veces, y aproximadamente una reducción de dos órdenes de magnitud en las necesidades de energía utilizando una variante personalizada de chips de procesamiento incrustados (Wehner et al., 2008).

15. Nordhaus (2007). Hay muchas interpretaciones de los diferentes significados de la ley de Moore; véase, por ejemplo, Tuomi (2002) y Mack (2011).
16. Si el desarrollo es lo suficientemente lento, el proyecto puede recurrir tanto a avances que surjan durante el proceso en el mundo exterior, como a avances en la ciencia de la computación realizados por investigadores universitarios y mejoras en el hardware realizados por la industria de los semiconductores.
17. El exceso algorítmico es tal vez menos probable, pero una excepción se daría si hardwares exóticos como la computación cuántica se volvieran capaces de ejecutar algoritmos que antes eran inviables. También se podría argumentar que las redes neuronales y el aprendizaje automático de profundidad son casos de exceso algorítmico: computacionalmente demasiado costosos como para funcionar bien cuando se inventaron por primera vez, fueron dejados de lado por un tiempo, y luego fueron desempolvados cuando las unidades de procesamiento gráfico rápido hicieron que fuera barato ejecutarlas. Ahora ganan concursos.
18. E incluso si el progreso en el camino hacia la línea de base humana fuera lento.
19. MundoD es la parte de optimización de energía del mundo que se aplicaría a la mejora del sistema en cuestión. Para un proyecto que funcionara en completo aislamiento, que no recibiera apoyo significativo permanente del mundo exterior, tendríamos un mundoD $\sim$ 0, a pesar de que el proyecto tendría que haber comenzado con una dotación de recursos (ordenadores, conceptos científicos, personal educado, etc.) que se derivarían de la economía mundial en su conjunto y de muchos siglos de desarrollo.
20. La más relevante de la habilidades cognitivas de la IA seminal en este caso es su capacidad para llevar a cabo diseños inteligentes para mejorarse a sí misma, es decir, su capacidad para amplificar su inteligencia. (Si la semilla IA es buena en la mejora de otro sistema, lo cual es bueno en la mejora de la semilla AI, entonces podríamos ver éstos como subsistemas de un sistema más amplio y centramos nuestro análisis en el todo mayor).
21. Esto supone que no se sabe que la resistencia al progreso sea tan alta como para desalentar a la inversión o para desviarla a algún proyecto alternativo.
22. Un ejemplo similar se discute en Yudkowsky (2008b).
23. Ya que las aportaciones han aumentado (por ejemplo, las inversiones en la construcción de nuevas fundiciones y el número de personas que trabajan en la industria de semiconductores), la ley de Moore en sí no ha aumentado rápidamente si tomamos en cuenta este aumento en las aportaciones. Combinado con los avances en software, sin embargo, un período de 18 meses de duplicación en el rendimiento por unidad aportada podría ser más plausible históricamente.
24. Se han llevado a cabo algunos intentos tentativos por desarrollar la idea de una explosión de inteligencia en el marco de la teoría del crecimiento económico; véase, por ejemplo, Hanson (1998b); Jones (2009); Salamon (2009). Estos estudios han señalado el potencial de crecimiento extremadamente rápido dada la llegada de las mentes digitales, pero puesto que la teoría del crecimiento endógeno está relativamente poco desarrollada incluso para aplicaciones históricas y contemporáneas, cualquier aplicación a un contexto futuro potencialmente discontinuo hay que entenderla más como una fuente de conceptos y consideraciones que como un ejercicio que pudiera ofrecernos de manera fiable previsiones autorizadas potencialmente útiles. Para una visión general de los intentos de modelar matemáticamente una singularidad tecnológica, véase Sandberg (2010).
25. Por supuesto, también es posible que no haya ningún despegue en absoluto. Mas, puesto que, como se argumentó anteriormente, la superinteligencia parece técnicamente factible, la ausencia de despegue sería debido probablemente a la intervención de algún impedimento, como una

catástrofe exis- tencial. Si llegara una superinteligencia fuerte no en la forma de inteligencia artificial o en la forma de emulación de cerebro completo, sino a través de uno de los otros caminos considerados antes, en ese caso un despegue lento sería más probable.

CAPÍTULO 5: VENTAJA ESTRATÉGICA DECISIVA

1. Una mente de software podría ejecutarse en una sola máquina en lugar de en una red mundial de computadoras; pero no es esto lo que entendemos por "concentración". En su lugar, lo que nos interesa aquí es el grado en que el poder, especialmente el poder derivado de la capacidad tecnológica, se concentrará en etapas avanzadas, o inmediatamente después, de la revolución en inteligencia artificial.
2. La difusión de tecnología como productos de consumo, por ejemplo, tiende a ser más lenta en los países en vías de desarrollo (Talukdar et al., 2002). Veáse también Keller (2004) y Banco Mundial (2008).
3. La literatura económica que se enfrenta a la teoría de la empresa es relevante como punto de comparación para la presente discusión. El lugar clásico es Coase (1937). Véase también, por ejemplo, Canbäck et al. (2006); Milgrom y Roberts (1990); Hart (2008); Simester y Knez (2002).
4. Por otro lado, podría ser especialmente fácil robar una IA seminal, ya que consistiría en software que podría ser transmitido electrónicamente o transportado en un dispositivo de memoria portátil.
5. Barber (1991) sugiere que la cultura Yangshao (5000-3000 a.C.) podría haber utilizado la seda. Sun et al. (2012) estiman, sobre la base de estudios genéticos, que la domesticación del gusano de seda habría ocurrido hace unos 4.100 años.
6. Cook (1984, 144). Esta historia podría ser demasiado buena como para resistir a un escrutinio histórico, como la historia de Procopio (Guerras VIII.xvii.1-7) de cómo los gusanos de seda fueron supuestamente llevados a Bizancio por monjes errantes, escondidos en sus bastones huecos de bambú (Hunt, 2011).
7. Wood (2007); Temple (1986).
8. Las culturas precolombinas tenían la rueda pero la utilizaban sólo como juguete (probablemente debido a la falta de buenos animales de tiro).
9. Koubi (1999); Lerner (1997); Koubi y Lalman (2007); Zeira (2011); Judd et al. (2012).
10. Estimado a partir de varias fuentes. El intervalo de tiempo es a menudo un tanto arbitrario, dependiendo de cómo definamos exactamente las capacidades "equivalentes". El radar fue utilizado por al menos dos países sólo un par de años después de su introducción, pero las cifras exactas en meses son difíciles de conseguir.
11. La RDS-6 en 1953 fue la primera prueba de bomba con reacciones de fusión, pero la RDS-37 en 1955 fue la primera "verdadera" bomba de fusión, donde la mayor parte de la energía provenía de la reacción de fusión.
12. Sin confirmar.
13. Pruebas en 1989, proyecto cancelado en 1994.
14. Sistema desplegado, con un alcance superior a 5.000 km.
15. Misiles Polaris comprados a los EE.UU.
16. El trabajo sobre los misiles Taimur está en marcha actualmente, probablemente basado en en misiles chinos.
17. Las pruebas del cohete RSA-3 de 1989 hasta 1990 estaban pensadas para lanzamientos de satélite y/o como un ICBM.
18. MIRV = vehículo de reentrada de blanco múltiple e independiente, una tecnología que permite que un solo misil balístico tenga múltiples ojivas que pueden ser programadas para golpear a objetivos diferentes.
19. El sistema Agni V aún no está disponible.
20. Ellis (1999).
21. Si modelamos la situación como una en la que el tiempo de espera entre los

proyectos se planee en base a una distribución normal, entonces la distancia probable entre el proyecto en cabeza y su seguidor más cercano también dependerá de la cantidad de proyectos que haya. Si hay un gran número de proyectos, entonces la distancia entre los dos primeros es probable que sea pequeña incluso si la variación de la distribución es moderadamente alta (aunque la diferencia esperada entre el primero y el segundo proyecto disminuya muy lentamente con el número de competidores si los plazos de ejecución se distribuyen normalmente). Sin embargo, es poco probable que haya un gran número de proyectos que se encuentren cada uno con recursos suficientemente buenos como para ser serios contendientes. (Puede haber un mayor número de proyectos si hubiera un gran número de diferentes enfoques básicos que pudieran ser perseguidos, pero en ese caso muchos de esos enfoques es probable que acabaran siendo callejones sin salida). Como se ha sugerido, empíricamente nos parece encontrar que no suele haber más de un puñado de competidores serios que persigan algún objetivo tecnológico específico. La situación es algo diferente en un mercado de consumo, donde hay muchos nichos para productos ligeramente diferentes y donde las barreras de entrada son bajas. Hay un montón de proyectos de una sola persona dedicados a diseñar camisetas, pero sólo unas pocas empresas en el mundo desarrollan la próxima generación de tarjetas gráficas. (Dos empresas, AMD y NVIDIA, disfrutan de un casi duopolio por el momento, aunque Intel también está compitiendo en el extremo inferior de rendimiento del mercado).

22. Bostrom (2006c). Uno podría imaginar una Unidad cuya existencia fuera invisible (por ejemplo, una superinteligencia con tecnología avanzada o intuición que sutilmente pudiera controlar los acontecimientos del mundo sin que ningún humano notara sus intervenciones); o una Unidad que se impusiera voluntariamente limitaciones muy estrictas en su ejercicio del poder (por ejemplo, limitándose puntillosamente a asegurar que ciertas reglas o tratados internacionales especificados —o principios libertarios— fueran respetados). Cuán probable sería que algún tipo particular de Unidad surgiera es, por supuesto, una cuestión empírica; pero conceptualmente, al menos, es posible tener una buena Unidad, una mala Unidad, una Unidad revoltosamente diferente, una Unidad de suavidad monolítica, una Unidad atenazante y opresiva, o una Unidad más parecida a una ley de la naturaleza que a un déspota vociferante.
23. Jones (1985, 344).
24. Puede ser significativo que el Proyecto Manhattan se llevara a cabo en tiempos de guerra. Muchos de los científicos que participaron afirmaron estar motivados principalmente por la situación de guerra y el temor a que la Alemania nazi pudiera desarrollar armas atómicas antes que los aliados. Podría ser difícil para muchos gobiernos movilizar un esfuerzo igualmente intenso y secreto en tiempos de paz. El programa Apolo, otro emblemático megaproyecto de ciencia/ingeniería, recibió un fuerte impulso de la rivalidad de la Guerra Fría.
25. Aunque incluso si buscaran arduamente, no está claro que fueran a mostrar (públicamente) que lo estaban haciendo.
26. Las técnicas criptográficas podrían permitir que el equipo de colaboradores estuviera separado físicamente. El único eslabón débil en la cadena de comunicación podría ser la etapa de entrada, donde potencialmente podría observarse el acto físico de escribir. Pero si la vigilancia en interiores se hiciera común (por medio de dispositivos de grabación microscópicos), los entusiastas de la protección de la privacidad podrían desarrollar contramedidas (por ejemplo, armarios especiales que pudieran sellarse al margen de posibles dispositivos de escucha). Independientemente de que el espacio físico pudiera llegar a ser transparente en una era de vigilancia futura, el ciberespacio podría posiblemente estar más protegido gracias a una mayor adopción de protocolos criptográficos más fuertes.
27. Un Estado totalitario podría recurrir a medidas aún más coercitivas. Los

científicos de campos relevantes podrían ser secuestrados y puestos en campos de trabajo, de manera similar a los "pueblos académicos" de la Rusia estalinista.

28. Cuando el nivel de preocupación pública fuera relativamente bajo, algunos investigadores podrían agradecer cierto alarmismo público que llame la atención sobre su trabajo y haga que el ámbito donde trabajen parezca importante y emocionante. Cuando el nivel de preocupación se elevara, las comunidades de investigación pertinentes podrían cambiar de tono a medida que comenzaran a preocuparse por los recortes de fondos, por la regulación y por reacción negativa del público. Los investigadores en disciplinas vecinas —como aquellas partes de la informática y la robótica que no son muy relevantes para la inteligencia artificial general— podrían resentirse de la huida de la financiación y la atención en sus propias áreas de investigación. Estos investigadores también podrían observar correctamente que su trabajo no conlleva ningún riesgo de llevar a una explosión de la inteligencia peligrosa. (Algunos paralelismos históricos podrían trazarse con el desarrollo de la idea de la nanotecnología, véase Drexler [2013]).
29. Éstos han tenido éxito en haber logrado al menos parte de lo que se propusieron hacer. Es difícil de determinar cómo de exitosos han sido en un sentido más amplio (tomando en cuenta la rentabilidad, etc.). En el caso de la Estación Espacial Internacional, por ejemplo, se han producido enormes sobrecostes y retrasos. Para los detalles de los problemas encontrados por el proyecto, véase NASA (2013). El proyecto del Gran Colisionador de Hadrones ha tenido algunos reveses importantes, pero esto podría ser debido a la dificultad inherente de la tarea. El Proyecto Genoma Humano alcanzó el éxito al final, pero parece haber recibido un aumento de velocidad al verse obligado a competir con el esfuerzo corporativo privado de Craig Venter. Proyectos patrocinados a nivel internacional para lograr la energía de fusión controlada no han logrado cumplir con las expectativas, a pesar de disponer de una inversión masiva; mas esto podría ser atribuible a que la tarea resultara ser más difícil de lo previsto.
30. Congreso de los EE.UU., Oficina de Evaluación Tecnológica (1995).
31. Hoffman (2009); Rhodes (2008).
32. Rhodes (1986).
33. La organización encargada de desbloquear códigos de la Armada de Estados Unidos, OP-20-G, ignoró, al parecer, una invitación para tener pleno conocimiento de los métodos anti-Enigma de Gran Bretaña, y no informó a los responsables de más alto nivel de Estados Unidos de la oferta de Gran Bretaña para compartir sus secretos criptográficos (Burke, 2001). Esto dio a los líderes estadounidenses la impresión de que Gran Bretaña estaba ocultando información importante, una causa de fricción durante toda la guerra. Gran Bretaña sí compartió con el gobierno soviético parte de los datos de inteligencia de comunicación alemana descifrados que habían conseguido. En particular, Rusia fue advertida acerca de los preparativos alemanes para la Operación Barbarroja. Pero Stalin se negó a creer la advertencia, en parte porque los británicos no dieron a conocer la forma en que habían obtenido la información.
34. Durante unos años, Russell parece haber abogado por la amenaza de una guerra nuclear sobre Rusia para que aceptaran el plan Baruch; más tarde, fue un fuerte defensor del desarme nuclear mutuo (Russell y Griffin, 2001). Se ha sabido que John von Neumann opinaba que una guerra entre Estados Unidos y Rusia era inevitable, y afirmó que: "Si preguntas por qué no bombardearlos [a los rusos] mañana, digo ¿por qué no bombardearlos hoy? Si dices hoy a las cinco de la tarde, yo digo ¿por qué no a la una?" (Es posible que él hiciera esta notoria declaración para pulir sus credenciales anticomunistas frente a los halcones estadounidense de Defensa de la era McCarthy. Si von Neumann, en caso de haber estado a cargo de la política estadounidense, hubiera llegado realmente a lanzar un primer ataque es

imposible saber a ciencia cierta. Véase Blair [1957], 96).

35. Baratta (2004).
36. Si la IA es controlada por un grupo de seres humanos, el problema puede aplicarse a ese grupo humano, aunque es posible que nuevas formas de llegar de forma fiable a un acuerdo estén disponibles para estas fechas, en cuyo caso incluso grupos humanos podrían evitar este problema de potencial desmoronamiento interno y derrocamiento mediante una sub-coalición.

CAPÍTULO 6: SUPERPODERES COGNITIVOS

1. ¿En qué sentido es la humanidad una especie dominante en la Tierra? Ecológicamente hablando, el ser humano es el animal grande (~ 50 kg) más común, pero la biomasa total humana seca (~ 100 millones de kg) no es tan impresionante en comparación con la de las hormigas, la familia Formicidae (300 miles de millones-3.000 miles de millones de kg). Los humanos y los organismos de servicios públicos humanos forman una parte muy pequeña (<0,001) de la biomasa total mundial. Sin embargo, las tierras de cultivo y pastizales son ahora uno de los mayores ecosistemas del planeta, que cubren aproximadamente el 35% de la superficie terrestre libre de hielo (Foley et al., 2007). Y de acuerdo con una evaluación típica (Haberl et al., 2007) nos apropiamos de casi un cuarto de la productividad primaria neta, aunque las estimaciones varían del 3 a más del 50%, dependiendo principalmente de la variación en las definiciones de los términos pertinentes (Haberl et al., 2013). Los seres humanos también tienen la mayor cobertura geográfica de las especies animales y encabezan el mayor número de cadenas diferentes de comida.
2. Zalasiewicz et al. (2008).
3. Véase la primera nota a este capítulo.
4. Estrictamente hablando, esto puede no ser del todo correcto. La inteligencia en la especie humana se extiende a lo largo de todos los rangos hasta aproximadamente cero (por ejemplo, en el caso de los embriones o de los pacientes en estado vegetativo permanente). En términos cualitativos, la máxima diferencia en la capacidad cognitiva dentro de la especie humana, por tanto, es tal vez mayor que la diferencia entre un humano y una superinteligencia. Pero el argumento del texto se mantiene si entendemos como "humano" un "adulto funcionalmente normal"
5. Gottfredson (2002). Véase también Carroll (1993) y Deary (2001).
6. Véase Legg (2008). A grandes rasgos, Legg propone medir un agente de aprendizaje por refuerzo como la rentabilidad esperada en todos los entornos de recompensa sumables, donde cada uno de estos ambientes recibe un peso determinado en función de su complejidad de Kolmogorov. Explicaremos qué se entiende por aprendizaje por refuerzo en el Capítulo 12. Véase también Dowe y Hernández-Orallo (2012) y Hibbard (2011).
7. Con respecto a la investigación en tecnología de áreas como la biotecnología y la nanotecnología, en lo que una superinteligencia podría sobresalir es en el diseño y modelado de nuevas estructuras. En la medida en que el ingenio de diseño y modelado no puedan sustituir a la experimentación física, la ventaja de rendimiento de la superinteligencia puede ser calificada por su nivel de acceso al necesario aparato experimental.
8. Por ejemplo, Drexler (1992, 2013).
9. Una IA de dominio restringido podría, por supuesto, tener aplicaciones comerciales importantes, pero esto no querría decir que tuviera el superpoder de la productividad económica. Por ejemplo, aunque una IA de dominio restringido ganara a sus propietarios varios miles de millones de dólares al año, esto seguiría siendo cuatro órdenes de magnitud menos que el resto de la economía mundial. Para que el sistema aumentara de manera directa y decisiva el producto mundial, una IA tendría que ser capaz de realizar muchos tipos de trabajo; es decir, tendría que tener competencia en

muchos ámbitos.

10. El criterio no excluye todos los escenarios en los que falla la IA. Por ejemplo, la IA puede racionalmente hacer una apuesta en la que tenga una alta probabilidad de perder. En este caso, sin embargo, el criterio podría tomar la siguiente forma: que (a) la IA debería estar haciendo una estimación desprejuiciada de las pocas probabilidades de éxito de la apuesta y (b) no debería haber una mejor apuesta a disposición de la IA que los seres humanos pudieran pensar pero que a la IA se le pasaría por alto.
11. Cf. Freitas (2000) y Vassar y Freitas (2006).
12. Yudkowsky (2008a).
13. Freitas (1980); Freitas y Merkle (2004, cap. 3); Armstrong y Sandberg (2013).
14. Véase, por ejemplo, Huffman y Pless (2.003), Knill et al. (2000), Drexler (1986).
15. Es decir, que la distancia sería pequeña en alguna métrica "natural", como el logaritmo del tamaño de la población que podría ser sostenible apoyado a nivel de subsistencia por un determinado nivel de capacidad si todos los recursos se dedicaran a ese fin.
16. Esta estimación se basa en la estimación de WMAP de una densidad cosmológica bariónica de 9,9 x 10-30 g/cm^3 y asume que el 90% de la masa es gas intergaláctico, que un 15% de la masa galáctica son estrellas (alrededor del 80% de materia bariónica), y que la estrella promedio pesa 0,7 masas solares (Lea y Trentham, 2005; Carroll y Ostlie, 2007).
17. Armstrong y Sandberg (2013).
18. Incluso a un 100% de c (que es inalcanzable para los objetos con masa en reposo mayor que cero) el número de galaxias alcanzables es sólo alrededor de 6 x 10^9. (Cfr Gott et al. [2005] y Heyl [2005]). Estamos asumiendo que nuestra comprensión actual de la física relevante es correcta. Es difícil tener ninguna confianza de nivel superior, ya que es al menos concebible que una civilización superinteligente pudiera extender su alcance de alguna manera que llegáramos a ser físicamente imposibles (por ejemplo, mediante la construcción de máquinas del tiempo, mediante el la producción de nuevos universos inflacionarios o mediante algunos otros medios aún inimaginables).
19. El número de planetas habitables por estrella es incierto, por lo que esto no es más que una estimación aproximada. Traub (2012) predice que un tercio de las estrellas en las clases espectrales F, G, K o tiene al menos un planeta terrestre en la zona habitable; véase también Clavin (2012). Las estrellas FGK forman aproximadamente el 22,7% de las estrellas de la vecindad solar, lo que sugiere que el 7,6% de las estrellas tienen planetas potencialmente adecuados. Además, puede haber planetas habitables alrededor de las estrellas M, que son más numerosas (Gilster, 2012). Véase también Robles et al. (2008). No sería necesario someter los cuerpos humanos a los rigores de los viajes intergalácticos. Una IA podría supervisar el proceso de colonización. Los *homo sapiens* podrían ser llevados como información, de tal modo que la IA podría más adelante utilizarla para crear ejemplares de nuestra especie. Por ejemplo, la información genética se podría sintetizar en ADN, y una primera generación de seres humanos podría ser incubada, criada, y educada por tutores IA que tomaran una apariencia antropomórfica.
20. O'Neill (1974).
21. Dyson (1960) afirma haber tomado la idea básica del escritor de ciencia ficción Olaf Stapledon (1937), que a su vez podría haberse inspirado por pensamientos similares de J.D. Bernal (Dyson, 1979, 211).
22. El principio de Landauer declara que hay una cantidad mínima de energía necesaria para cambiar un bit de información, conocida como el límite de Landauer, igual a kT ln 2, donde k es la constante de Boltzmann (1,38 x 10^{-23} J/K) y T es la temperatura. Si asumimos que el circuito se mantiene a alrededor de 300 K, entonces 10^{26} vatios nos permitirían borrar

aproximadamente 10^{47} bits por segundo. (Para la eficiencia alcanzable por dispositivos computacionales nanomecánicos, véase Drexler [1992]. Véase también Bradbury [1999]; Sandberg [1999]; Cirkovic [2004]. Los fundamentos del principio de Lan- dauer son todavía algo que se discute; véase, por ejemplo, Norton [2011]).

23. Las estrellas varían en su energía proporcionada, pero el sol es una estrella de secuencia principal bastante típica.
24. Un análisis más detallado podría considerar mejor qué tipo de cálculo nos interesa. El número de cálculos de serie que se pueden realizar es bastante limitado, ya que un ordenador serial rápido debe ser pequeño para reducir al mínimo los retardos comunicativos entre las diferentes partes de la computadora. También hay límites en el número de bits que se pueden almacenar, y, como hemos visto, en el número de pasos de cálculo irreversibles (que implican el borrado de información) que se pueden realizar.
25. Estamos asumiendo aquí que no hay civilizaciones extraterrestres que puedan interponerse en el camino. También estamos asumiendo que la hipótesis de la simulación es falsa. Veáse Bostrom (2003a). Si cualquiera de estos supuestos fuera incorrecto, podría haber riesgos no antropogénicos importantes —aquellos que impliquen la agencia inteligente de especies no humanas. Veáse también Bostrom (2003b, 2009c).
26. Al menos una Unidad sabia que captara la idea de evolución podría, en principio, embarcarse en un programa eugenésico por medio del cual poco a poco podría elevar su nivel de inteligencia colectiva.
27. Tetlock y Belkin (1996).
28. Para que quede claro: la colonización y la manipulación técnica de una gran parte del universo accesible no está a nuestro alcance directo. La colonización intergaláctica está muy lejos del alcance de la tecno- logia de hoy. La cuestión es que podríamos utilizar, en principio, nuestras capacidades actuales para desarrollar las capacidades adicionales que serían necesarias, poniendo así el logro a nuestro alcance de manera indirecta. Por supuesto, es también cierto que la humanidad no es actualmente una Unidad y que no sabemos si nunca nos enfrentaremos a una oposición inteligente por parte de algún poder externo si comenzamos a rediseñar el universo accesible. Para cumplir con el umbral de sostenibilidad de Unidad-sabia, sin embargo, sería suficiente con poseer un conjunto de capacidades tales que, si una Unidad sabia que no tuviera oposición inteligente las hubiera poseído, entonces la colonización y el rediseño de una gran parte del universo accesible podría estar a su alcance indirecto.
29. A veces puede ser útil hablar de dos IAs como si cada una tuviera un superpoder distinto. En un sentido amplio de la palabra, uno podría por lo tanto concebir un superpoder como algo que un agente tiene en relación con algún campo de acción —en este caso, el campo tal vez incluya a toda la civilización humana, pero excluye a las otras IAs.

CAPÍTULO 7: LA VOLUNTAD SUPERINTELIGENTE

1. Esto por supuesto no pretende negar que las diferencias que parecen visualmente pequeñas pueden ser funcionalmente profundas.
2. Yudkowsky (2008a, 310).
3. David Hume, el filósofo escocés de la Ilustración, pensó que las creencias por sí solas (por ejemplo, las creencias sobre lo que es bueno hacer) no pueden motivar la acción: es necesario algún deseo. Esto apoyaría la tesis de ortogonalidad socavando una posible objeción a la misma, a saber, que una inteligencia suficiente podría implicar la adquisición de ciertas creencias que luego producirán necesariamente ciertas motivaciones. Sin embargo, aunque la tesis de ortogonalidad puede apoyarse en la teoría de la motivación de Hume, no la presupone. En particular, uno no necesita sostener que las creencias nunca pueden motivar por sí solas la acción. Sería suficiente asumir, por ejemplo, que un agente —que llegara a ser tan inteligente— podría motivarse a seguir cualquier curso de acción si el agente pasara

a tener ciertos deseos de suficiente fuerza primordial. Otra forma en que la tesis de ortogonalidad podría ser cierta incluso si la teoría de Hume de la motivación fuera falsa, es si una inteligencia arbitrariamente alta no implicara la adquisición de creencias (supuestamente) motivantes por su cuenta. Una tercera forma en que la tesis de la ortogonalidad podría ser cierta incluso si la teoría de Hume fuera falsa es si fuera posible la construcción de un agente (o de manera más neutral, un "proceso de optimización") con una inteligencia arbitrariamente grande pero con una constitución tan ajena que no tuviera analogías funcionales claras a lo que los seres humanos llamamos "creencias" y "deseos". (Para algunos intentos recientes de defender la teoría de Hume de la motivación véase Smith [1987], Lewis [1988], y Sinhababu [2009]).

4. Por ejemplo, Derek Parfit ha argumentado que ciertas preferencias básicas serían irracionales, como el que un agente, por lo demás normal, que tuviera "indiferencia respecto del próximo martes":

   Un cierto hedonista se preocupa mucho por la calidad de sus futuras experiencias. Con una excepción, se preocupa por igual de todas las partes de su futuro. La excepción es que tiene "indiferencia respecto del próximo martes". A lo largo de todos los martes se preocupa de manera normal de lo que le sucede. Pero nunca se preocupa por posibles dolores o placeres de un martes futuro... Esta indiferencia es un mero hecho. Cuando él está planeando su futuro, es simplemente verdad que siempre prefiere la perspectiva de un gran sufrimiento en un martes a un dolor leve cualquier otro día (Parfit [1986, 123-4]; véase también Parfit [2011]).

   Para nuestros propósitos, no necesitamos tomar ninguna postura sobre si Parfit está en lo cierto al decir que este agente es irracional, siempre y cuando aceptemos que no es necesariamente no-inteligente en el sentido instrumental que se explica en el texto. El agente de Parfit podría tener una racionalidad instrumental impecable, y por lo tanto una gran inteligencia, aunque se queda corto en algún tipo de sensibilidad hacia la "razón objetiva" que podría requerirse para ser un agente totalmente racional. Por lo tanto, este tipo de ejemplo no mina la tesis ortogonalidad.

5. Incluso si hubiera hechos morales objetivos que cualquier agente totalmente racional pudiera comprender, e incluso si estos hechos morales de alguna manera estuvieran intrínsecamente motivados (de tal manera que cualquiera que los comprendiera totalmente estaría necesariamente motivado a actuar de acuerdo con ellos), esta necesidad no socavaría la tesis de ortogonalidad. La tesis aún podría ser cierta si un agente pudiera tener una racionalidad instrumental impecable, incluso aunque careciera de alguna otra facultad constitutiva de una racionalidad adecuada, o alguna facultad necesaria para la plena comprensión de los hechos morales objetivos. (Un agente también podría ser muy inteligente, incluso superinteligente, sin tener una racionalidad instrumental plena en cada dominio).
6. Para más información sobre la tesis ortogonalidad, véase Bostrom (2012) y Armstrong (2013).
7. Sandberg y Bostrom (2008).
8. Stephen Omohundro ha escrito dos trabajos pioneros sobre este tema (Omohundro, 2007, 2008). Omohundro argumenta que todos los sistemas avanzados de IA son propensos a exhibir un número de "unidades básicas", que quiere decir "tendencias que estarán presentes si no son contrarrestadas de forma explícita" El término "unidad de IA" tiene la ventaja de ser breve y sugerente, pero tiene la desventaja de sugerir que los objetivos instrumentales a los que se refiere influyen en la toma de decisiones de la IA de la misma manera que las unidades psicológicas influyen en la toma de decisiones humanas, es decir, a través de una especie de un tirón fenomenológico en nuestro ego, que nuestra fuerza de voluntad puede en ocasiones resistir exitosamente. Esa connotación es inútil. Uno normalmente no diría que un ser humano típico tiene una "unidad" destinada a llenar su declaración de impuestos, a pesar de que la presentación de impuestos puede ser un objetivo fundamental bastante convergente para los seres humanos en las sociedades contemporáneas (un objetivo cuya realización evita problemas que nos impedirían realizar muchos de nuestros objetivos finales). Nuestro tratamiento aquí también difiere del de Omohundro en algunos temas más sustanciales, aunque la idea

subyacente es la misma. (Véase también Chalmers [2010] y Omohundro [2012]).

9. Chislenko (1997).
10. Véase también Shulman (2010b).
11. Un agente también podría cambiar su *representación* de objetivos si cambia su ontología, con el fin de adaptar su antigua representación a la nueva ontología; cf. de Blanc (2011). Otro factor que podría determinar que un teórico de la decisión evidencial emprendiera diversas acciones, incluyendo el cambio de sus objetivos finales, es la importación evidencial de la decisión de hacerlo. Por ejemplo, un agente que siguiera la teoría de la decisión evidencial podría creer que existen otros agentes como él en el universo y que sus propias acciones proporcionarían alguna evidencia acerca de cómo actuarán los otros agentes. Por lo tanto, el agente podría optar por adoptar un objetivo final que fuera altruista respecto de esos otros agentes evidencialmente vinculados, por la razón de que esto le daría al agente la evidencia de que esos otros agentes habrían optado por actuar de igual manera. Un resultado equivalente se podría obtener, sin embargo, sin cambiar los propios objetivos finales, eligiendo en cada instante actuar como si se tuviera esos objetivos finales.
12. Una extensa literatura psicológica explora la formación de preferencias adaptativas. Véase, por ejemplo, Forgas et al. (2010).
13. En los modelos formales, el valor de la información se cuantifica como la diferencia entre el valor esperado realizado por decisiones óptimas hechas con esa información y el valor esperado realizado por decisiones óptimas hechas sin ella. (Véase, por ejemplo, Russell y Norvig [2010]). De ello se deduce que el valor de la información nunca es negativo. También se deduce que cualquier información que usted sepa que nunca va a afectar a ninguna decisión que tendrá que realizar tiene un valor de cero para usted. Sin embargo, este tipo de modelo asume varias idealizaciones que a menudo no son válidas en el mundo real, tales como que el conocimiento no tiene valor final (lo que significa que el conocimiento sólo tiene valor instrumental y no tiene valor por sí mismo) y que los agentes no son transparentes para otros agentes.
14. Por ejemplo, Hájek (2009).
15. Esta estrategia está ejemplificada por la ascidia, un animal que nada hasta que encuentra una roca adecuada, a la que se fija permanentemente. Cimentado en su lugar, la ascidia tiene menos necesidad de un procesamiento complejo de la información, con lo que procede a digerir parte de su propio cerebro (el ganglio cerebral). Se puede observar el mismo fenómeno en algunos académicos cuando se les hace numerarios.
16. Bostrom (2012).
17. Bostrom (2006c).
18. Se podría revertir la cuestión y buscar en su lugar las posibles razones para que una Unidad superin- teligente no desarrollara algunas capacidades tecnológicas. Éstas incluyen las siguientes: (a) la Unidad prevé que no tendrá ningún uso para la capacidad; (b) el coste de desarrollo es demasiado grande en relación con su utilidad esperada (por ejemplo, si la tecnología no fuera adecuada para la consecución de cualquiera de los objetivos de la Unidad, o si la Unidad tuviera una tasa de ahorro muy alta que desalentara fuertemente la inversión); (c) la Unidad tiene algún valor final que requiere la abstención de vías particulares de desarrollo de la tecnología; (d) si la Unidad no está seguro de que se mantendrá estable, puede ser que prefiera abstenerse de tecnologías en desarrollo que podrían amenazar su estabilidad interna o que harían que las consecuencias de la disolución fuera peor (por ejemplo, un gobierno mundial puede no desear desarrollar tecnologías que facilitaran la rebelión, incluso si tuviera algunos usos buenos, ni desarrollar tecnologías que facilitaran la producción de armas de destrucción masiva que podrían causar estragos si el gobierno mundial se disolviera); (e) del mismo modo, la Unidad podría haber hecho algún tipo de compromiso de unión estratégica para no desarrollar una tecnología, un compromiso que seguiría vigente aunque ahora fuera conveniente desarrollarla. (Téngase en cuenta, sin embargo, que algunas de las razones actuales para el desarrollo de tecnología no se aplicarían a una Unidad: por ejemplo, las razones derivadas de las carreras armamentísticas).
19. Supongamos que un agente desprecia los recursos que podrían obtenerse en el futuro a un ritmo exponencial, y que debido a la limitación de la velocidad de la luz el

agente sólo pudiera aumentar su dotación de recursos a un ritmo polinómico. ¿Significaría esto que habría un momento después del cual al agente no le resultaría valioso continuar con la expansión adquisitiva? No, porque aunque el valor actual de los recursos obtenibles en ocasiones futuras fuera asíntota de cero cuanto más tendiéramos hacia el futuro, *también lo sería su coste actual de obtención.* El coste actual de envío de otra sonda von Neumann a más de 100 millones de años a partir de ahora (posiblemente mediante algún recurso adquirido poco tiempo antes) se vería reducido por el mismo factor de descuento que disminuiría el valor actual de los recursos futuros que la sonda adicional adquiriría (modulado como factor constante).

20. Mientras que el volumen alcanzado por las sondas de colonización en un momento dado puede ser más o menos esférico y en expansión con una velocidad proporcional al cuadrado del tiempo transcurrido desde la primera sonda que se puso en marcha (~ t2), la cantidad de recursos contenidos dentro de este volumen seguirá un patrón menor de crecimiento regular, ya que la distribución de recursos no es homogénea y varía a lo largo de varias escalas. Inicialmente, la tasa de crecimiento podría ser ~ t2 mientras el planeta de origen está siendo colonizado; entonces la tasa de crecimiento podría repuntar a medida que los planetas y sistemas solares cercanos fueran colonizados; entonces, a medida que se llenara el volumen más o menos en forma de disco de la Vía Láctea, la tasa de crecimiento podría estabilizarse, hasta ser aproximadamente proporcional a t; entonces la tasa de crecimiento de nuevo podría repuntar a medida que las galaxias cercanas fueran colonizadas; entonces la tasa de crecimiento podría volver a aproximarse a ~ t2 mientras la expansión continúa a una escala en la que la distribución de las galaxias es más o menos homogénea; luego otro período de repunte, seguido de un suave crecimiento de ~ t2 cuando los supercúmulos galácticos sean colonizados; hasta que en última instancia, la tasa de crecimiento inicie una disminución final, llegando finalmente a cero a medida que la velocidad de expansión del universo aumente hasta tal punto como para hacer que una ulterior colonización fuera imposible.
21. El argumento de la simulación puede ser de particular importancia en este contexto. Un agente super- inteligente puede asignar una probabilidad significativa a la hipótesis de que vive en una simulación por ordenador y su secuencia de percepciones es generada por otra superinteligencia, y esto podría generar diversas razones instrumentales convergentes dependientes de las conjeturas que el agente hiciera sobre la probabilidad de los distintos tipos de simulaciones en los que podría estar. Cf. Bostrom (2003a).
22. El descubrimiento de las leyes básicas de la física y otros hechos fundamentales sobre el mundo es un objetivo fundamental convergente. Podríamos colocarlo bajo la rúbrica de “mejora cognitiva”, aunque también podría ser derivado del objetivo de “perfección tecnológica” (ya que los fenómenos físicos nuevos podrían permitir tecnologías novedosas).

CAPÍTULO 8: ¿ES EL APOCALIPSIS EL RESULTADO INEVITABLE?

1. Existen algunos riesgos existenciales adicionales en escenarios en los que la humanidad sobrevive en un estado altamente subóptimo o en el que una gran parte de nuestro potencial para un desarrollo deseable se malgasta irreversiblemente. Además, puede haber riesgos existenciales asociados al período previo a una posible explosión de inteligencia, riesgos que surgirían, por ejemplo, de la guerra entre países que compiten por desarrollar primero la superinteligencia.
2. Habrá un momento importante de vulnerabilidad cuando la IA se dé cuenta por primera vez de la necesidad de dicho ocultamiento (un evento que podemos denominar el *origen del engaño).* Esta toma de conciencia inicial no se ocultaría deliberadamente cuando se produjera. Pero después de haberse dado cuenta, la IA podría moverse con rapidez para ocultar el hecho de que se hubiera producido tal descubrimiento, mientras crea alguna dinámica interna encubierta (tal vez disfrazada de proceso inocuo que se mezcla con el resto de procesos complicados que tienen lugar en su mente) que le

permitiría seguir planificando su estrategia a largo plazo en la intimidad.

3. Incluso los hackers humanos pueden escribir programas pequeños y aparentemente inocuos que hacen cosas totalmente inesperadas. (Para ejemplos, ver algunas de las obras ganadoras en el concurso Inter- national Obfuscated C Code).
4. El argumento de que algunas medidas de control de la IA podrían parecer trabajar dentro de un contexto fijo y, sin embargo, fallar catastróficamente cuando el contexto cambie, también fue destacado por Eliezer Yudkowsky; véase, por ejemplo, Yudkowsky (2008a).
5. El término parece haber sido acuñado por el escritor de ciencia ficción Larry Niven (1973), pero se basa en experimentos reales de estimulación cerebral de recompensa; cf. Olds y Milner (1954) y Oshima y Katayama (2010). Veáse también Ring y Orseau (2011).
6. Bostrom (1997).
7. Podría haber implementaciones del mecanismo de aprendizaje por refuerzo que, cuando la IA descubriera la solución de pinchazo cerebral, condujeran a una incapacitación segura en lugar de a una profusión infraestructural. El punto es que esto podría fácilmente torcerse y fracasar por razones inesperadas.
8. Esto fue sugerido por Marvin Minsky *(vide* Russell y Norvig [2010, 1039]).
9. La cuestión de qué tipo de mentes digitales serían conscientes, en el sentido de tener experiencia fenoménica subjetiva, o "qualia" en lenguaje filósofico, es importante en relación a este punto (aunque es irrelevante para muchas otras partes de este libro). Una pregunta abierta es cómo de difícil sería estimar con precisión cómo se comportaría un ser humano en varias circunstancias sin estar simulando su cerebro con detalle suficiente como para que la simulación fuera consciente. Otra cuestión es si hay en general algoritmos útiles para una superinteligencia, por ejemplo técnicas de aprendizaje por refuerzo, de tal forma que la implementación de estos algoritmos generaría qualia. Incluso si juzgamos la probabilidad como muy improbable que cualquiera de estos subprogramas fuera consciente, el número de instancias puede ser tan grande que incluso un pequeño riesgo de que puedan experimentar sufrimiento debe ser muy tenido en cuenta en nuestro cálculo moral. Véase también Metzinger (2003, cap. 8).
10. Bostrom (2002a, 2003a); Elga (2004).

CAPÍTULO 9: EL PROBLEMA DEL CONTROL

1. Por ejemplo, Laffont y Martimort (2002).
2. Supongamos que una mayoría de votantes quiere que su país construya algún tipo particular de su- perinteligencia. Ellos eligen a un candidato que promete cumplir sus órdenes, aunque podrían tener dificultades para asegurarse de que el candidato, una vez en el poder, siga adelante con su promesa de campaña y lleve a cabo el proyecto en la forma en que los votantes querían. Suponiendo que fuera fiel a su palabra, el gobernante instruiría a su gobierno a contratar un consorcio académico o una industria para que llevara a cabo el trabajo; pero de nuevo habría problemas de agencia: los burócratas en el departamento de gobierno podrían tener sus propios puntos de vista sobre lo que debe hacerse y podrían poner en práctica el proyecto de una manera que respete la letra pero no el espíritu de las instrucciones del líder. Incluso si el departamento del gobierno hace su trabajo fielmente, los socios científicos contratados podrían tener sus propias intenciones distintas. El problema se repite a muchos niveles. El director de uno de los laboratorios participantes podría tener la preocupación de que un técnico que introdujera un elemento no autorizado en el diseño —imaginando que el Dr. T.R. Aición se cuela en su oficina a última hora de la noche, se registra en el código base del proyecto, y reescribe una parte de la IA seminal del sistema objetivo. Donde se supone que debería poner "servir a la humanidad", ahora diría "servir al Dr. T.R. Aición."

3. Incluso para el desarrollo de la superinteligencia podría haber lugar para pruebas de comportamiento —como elemento auxiliar dentro de una batería más amplia de medidas de seguridad. En caso de que una IA se portarse mal en su fase de desarrollo, algo estaría claramente estropeado— aunque, de manera importante, la inversa no se cumple.
4. En un pirateo clásico de 1975, Steven Dompier escribió un programa para la Altair 8800 que se aprovechó de este efecto (y de la ausencia de protección alrededor de la caja de la microcomputadora). La ejecución del programa provocó la emisión de ondas electromagnéticas que producían música cuando se acercaba una radio al ordenador (Driscoll, 2012). El joven Bill Gates, que asistió a una demostración, informó de que estaba impresionado y perplejo por el pirateo (Gates, 1975). Hay, en todo caso, los planes para diseñar futuros chips con capacidades integradas de Wi-Fi (Greene, 2012).
5. No es una cuestión baladí haber mantenido una convicción que, de haber tenido la oportunidad de actuar en consecuencia, podría haber acabado arruinando todos nuestros cósmicos. Tal vez podría enunciarse el siguiente principio: si alguien en el pasado ha determinado en N ocasiones que un sistema ha sido mejorado lo suficiente como para que sea seguro, y cada vez se supo que estaba equivocado, entonces en la siguiente ocasión no tiene derecho a que se asigne un crédito mayor a $1/(N+1)$ a la afirmación de que el sistema es seguro.
6. En un experimento informal, el papel de la IA fue interpretado por un humano inteligente. Otro individuo desempeñó el papel de guardián y tuvo la tarea de no dejar que la IA saliera de la caja. La IA podía comunicarse con el guardián solamente por texto y se le dio dos horas para persuadir al guardián de que le dejara salir. En tres casos de cada cinco, con diferentes individuos interpretando al guardián, la IA se escapó (Yudkowsky, 2002). Lo que un humano puede hacer, una superinteligencia también podría. (Lo contrario, por supuesto, no se sostiene. Incluso si la tarea para una superinteligencia real fuera más difícil —tal vez los guardianes estarían más fuertemente motivados para abstenerse de liberar a la IA que los individuos que hacían de guardián en el experimento— la superinteligencia aún podría tener éxito donde un humano fracasaría).
7. No deberíamos exagerar la cantidad marginal de seguridad que podría obtenerse de esta manera. Las imágenes mentales pueden sustituir a la pantalla gráfica. Considérese el impacto que pueden tener los libros en las personas —y los libros ni siquiera son interactivos.
8. Veáse también Chalmers (2010). Sería un error deducir de esto que no hay uso posible en la construcción de un sistema que nunca sería observado por cualquier entidad externa. Se podría dar un valor final a lo que sucede en el interior de un sistema de este tipo. Además, otras personas podrían tener preferencias acerca de lo que sucede en el interior de un sistema de este tipo, y podrían, por lo tanto, estar influenciados por su creación o por la promesa de su creación. El conocimiento de la existencia de ciertos tipos de sistemas aislados (los que contienen observadores) también puede inducir cierta incertidumbre antrópica en observadores externos, lo cual podría influir en su comportamiento.
9. Uno podría preguntarse por qué la integración social se considera una forma de control de capacidad. ¿No debería clasificarse más bien como un método de selección de la motivación en base a que implica tratar de influir en el comportamiento de un sistema por medio de incentivos? Ahora abordaremos detalladamente la selección de motivación; pero, en respuesta a esta pregunta, estamos interpretando la selección de motivación como un conjunto de métodos de control que funcionan mediante la selección o la conformación de los objetivos finales de un sistema —objetivos buscados por sí mismos y no por razones instrumentales. La integración social no se dirige a los objetivos finales de un sistema, por lo que no es una selección por motivación. Más bien, la integración social tiene como objetivo limitar

las capacidades efectivas del sistema: se trata de hacer que el sistema sea incapaz de lograr un cierto conjunto de resultados—, resultados en los que el sistema alcanza los beneficios de la deserción sin sufrir las penas asociadas (retribución y pérdida de las ganancias de la colaboración). La esperanza es que mediante la limitación de que los resultados que el sistema es capaz de alcanzar, el sistema encontraría que los medios restantes más eficaces para la realización de sus objetivos finales sea comportarse de manera cooperativa.

10. Este enfoque podría ser algo más prometedor en el caso de una emulación que creyéramos que tuviera motivaciones antropomórficas.
11. Debo esta idea a Carl Shulman.
12. Crear un sistema de cifrado seguro para resistir a un descifrador superinteligente es un desafío nada trivial. Por ejemplo, los rastros de números aleatorios pueden ser puestos en el cerebro de algún observador o en la microestructura del generador aleatorio, de donde la superinteligencia puede recuperarlos; o, si se utilizan números pseudo-aleatorios, la superinteligencia podría adivinarlos o descubrir la semilla de la que surgieron. Además, la superinteligencia podría construir enormes ordenadores cuánticos, o incluso descubrir fenómenos físicos desconocidos que podrían ser utilizados para construir nuevos tipos de ordenadores.
13. La IA podría pincharse a sí misma para creer que había recibido unas fichas de recompensa, pero esto no haría que estuviera pirateada si estaba diseñada para querer fichas de recompensa (en lugar de querer estar en un estado en el que tuviera ciertas creencias acerca de las fichas de recompensa).
14. Para el artículo original, consulte Bostrom (2003a). Véase también Elga (2004).
15. Shulman (2010a).
16. La realidad a nivel de base probablemente contiene más recursos computacionales que la realidad simulada, ya que todos los procesos computacionales que ocurren en una simulación también se están produciendo en el equipo en el que se ejecuta la simulación. La realidad a nivel de sótano también podría contener una gran cantidad de otros recursos físicos a los que podrían ser difícil de acceder para los agentes simulados —agentes que sólo existen por la indulgencia de simuladores de gran alcance que podrían tener otros usos en mente para esos recursos. (Por supuesto, la inferencia aquí no es estrictamente válida de manera deductiva: en principio, podría darse el caso de que los universos en los que se ejecutaran las simulaciones contuvieran muchos más recursos que los que tienen las civilizaciones no simuladas —incluso aunque cada civilización no simulada que ejecutara simulaciones tuviera más recursos que todas las civilizaciones simuladas combinadas).
17. Hay varias consideraciones esotéricas adicionales que podrían importar en esta materia, las implicaciones de las cuales aún no han sido plenamente analizadas. Estas consideraciones pueden en última instancia ser de vital importancia en el desarrollo de un enfoque exhaustivo para hacer frente a la perspectiva de una explosión de inteligencia. Sin embargo, parece poco probable que vayamos a tener éxito en averiguar la importancia práctica de tales argumentos esotéricos a menos que primero hagamos cierto progreso en consideraciones más mundanas que se abordan en la mayor parte de este libro.
18. Cf., por ejemplo, Quine y Ullian (1978).
19. Los cuales podrían ser investigados por una IA considerando las características de rendimiento de varias funcionalidades computacionales básicas, tales como el tamaño y la capacidad de los diversos bloques de datos, el tiempo que se necesita para acceder a diferentes partes de la memoria, la incidencia de los giros de bits aleatorios, y así sucesivamente.
20. Tal vez el principio podría ser (una aproximación computable de) el principio de Solomonoff, que asigna una probabilidad para los mundos

posibles en base a su complejidad algorítmica. Véase Li y Vitányi (2008).

21. En el momento después del origen del engaño, la IA podría ingeniárselas para borrar el rastro de su pensamiento rebelde. Por tanto, es importante que el cable trampa opere continuamente. También sería una buena práctica utilizar un "registrador de vuelo" que almacene un registro completo de toda la actividad de la IA (incluyendo la fecha exacta de entradas de teclado de los programadores), por lo que su trayectoria pudiera ser desandada o analizada después de una parada automática. La información podría ser almacenada en un medio que sólo pudiera escribirse una vez pero leerse muchas.
22. Asimov (1942). A las tres leyes fueron posteriormente se les añadió una "Ley Cero": "(0) Un robot no debe dañar a la humanidad o, por inacción, permitir que la humanidad sufra daño" (Asimov, 1985).
23. Cf. Gunn (1982).
24. Russell (1986, 161f).
25. Del mismo modo, aunque algunos filósofos han empleado carreras enteras en tratar de formular cuidadosamente sistemas deontológicos, de vez en cuando salen a la luz nuevos casos y consecuencias que requieren revisiones. Por ejemplo, la filosofía moral deontológica se ha fortalecido en los últimos años a través del descubrimiento de una nueva clase fértil de experimentos filosóficos de pensamiento, los "problemas de la vagoneta", que revelan muchas interacciones sutiles entre nuestras intuiciones sobre el significado moral de la distinción entre actos/omisiones, la distinción entre consecuencias deseadas e indeseadas, y otras cuestiones; véase, por ejemplo, Kamm (2007).
26. Armstrong (2010).
27. Como regla general, si uno planea utilizar varios mecanismos de seguridad para contener una IA, puede ser sabio trabajar en cada uno como si estuviera destinado a ser el único mecanismo de seguridad y como si estuviera, por tanto, obligado a ser suficiente de manera individual. Si uno pusiera un cubo agujereado dentro de otro cubo agujereado, el agua seguiría saliéndose.
28. Una variación de la misma idea es construir la IA para que estuviera continuamente motivada a actuar en base a sus mejores conjeturas sobre el estándar implícitamente definido. En esta configuración, el objetivo final de la IA es siempre actuar según el estándar implícitamente definido, y proseguir una investigación sobre esta norma sólo por razones instrumentales.

CAPÍTULO 10: ORÁCULOS, GENIOS, SOBERANOS, HERRAMIENTAS

1. Estos nombres son, evidentemente, antropomorfos y no debe ser tomados en serio como analogías. Están pensados sólo como etiquetas para algunos conceptos diferentes prima facie de posibles tipos de sistema que podrían tratarse de construir.
2. En respuesta a una pregunta sobre el resultado de las próximas elecciones, uno no desea ser abrumado con una lista completa de la posición y el momento del vector proyectado de las partículas cercanas.
3. Indicado a seguir un conjunto de instrucciones particulares en una máquina particular.
4. Kuhn (1962); de Blanc (2011).
5. Sería más difícil aplicar un "método de consenso" a genios o soberanos, ya que a menudo puede haber numerosas secuencias de acciones básicas (como el envío de patrones particulares de señales eléctricas a los mecanismos del sistema) que serían casi exactamente igual de eficaces a la hora de lograr un objetivo determinado; mientras que agentes ligeramente diferentes podrían legítimamente elegir acciones ligeramente diferentes, lo que resultaría un fracaso en llegar a un consenso. En contraste, con preguntas adecuadamente formuladas habría, por lo general, un pequeño número de opciones de respuesta adecuadas (tales como "sí" y "no"). (Sobre el concepto de punto de Schelling, también referido como "punto focal", véase Schelling [1980]).

6. ¿No es la economía mundial, en algunos aspectos, análoga a un genio débil, si bien uno que cobra por sus servicios? Una economía mucho más grande, tal como podría desarrollarse en el futuro, podría aproximarse a un genio con superinteligencia colectiva.

   Un aspecto importante en el que la economía actual *no se parece* a un genio es que si bien puedo ordenar a la economía que mande una pizza a mi domicilio (pagando), no puedo ordenarle que me mande la paz mundial. La razón no es que la economía no sea lo suficientemente potente, sino que está insuficientemente coordinada. En este sentido, la economía se asemeja a una *asamblea* de genios que sirven a diferentes maestros (con objetivos enfrentados) más de lo que se asemeja a un solo genio o a cualquier otro tipo de agente unificado. El aumento de la potencia total de la economía al hacer que cada genio constituyente sea más potente, o añadiendo más genios, no llevaría necesariamente a que la economía fuera más capaz de proporcionar la paz. Para funcionar como un genio superinteligente, la economía no sólo tendría que crecer en su capacidad de producir bienes y servicios a bajo coste (incluyendo aquellos que requieren tecnología radicalmente nueva), también necesitaría estar en mejores condiciones de solucionar los problemas de coordinación mundial.
7. Si el genio fuera de alguna manera incapaz de no obedecer una orden —y de alguna manera incapaz de reprogramarse a sí mismo para deshacerse de esta condición— entonces podría actuar para prevenir que cualquier nueva orden fuera dada.
8. Incluso un oráculo que se limitara a dar respuestas sí/no podría utilizarse para facilitar la búsqueda de un genio o de un soberano IA, o incluso podría ser utilizado directamente como un componente de una IA de este tipo. El oráculo también podría ser utilizado para producir el código real de una IA si se le pudiera hacer un número suficientemente grande de preguntas. Una serie tal de preguntas podría tomar más o menos la siguiente forma: “En la versión binaria del código de la primera IA que usted pensó que constituiría a un genio, ¿es el símbolo X un cero?”.
9. Uno podría imaginar un oráculo un poco más complicado o un genio que aceptara preguntas u órdenes sólo si fueran emitidas por una autoridad designada, aunque esto todavía dejaría abierta la posibilidad de que la autoridad se corrompiera o de que fuera chantajeado por un tercero.
10. John Rawls, prominente filósofo político del siglo XX, empleó el famoso dispositivo expositivo del velo de ignorancia como forma de caracterizar el tipo de preferencia que debería tenerse en cuenta en la formulación de un contrato social. Rawls sugiere que deberíamos imaginarnos eligiendo un contrato social detrás de un velo de ignorancia que nos impide saber qué persona seremos y qué rol social ocuparemos, con la idea de que en una situación así tendríamos que pensar qué sociedad sería generalmente más justa y más deseable sin tener en cuenta nuestros intereses egoístas y prejuicios egoístas que podrían hacernos preferir un orden social en el que nosotros mismos gozaríamos de privilegios injustos. Veáse Rawls (1971).
11. Karnofsky (2012).
12. Una posible excepción sería el software conectado a mecanismos suficientemente potentes, como si se conectara el software de los sistemas de alerta temprana a las cabezas nucleares o a los oficiales humanos autorizados para lanzar un ataque nuclear. Fallos en este tipo de software pueden dar lugar a situaciones de alto riesgo. Esto ha sucedido por lo menos dos veces en la historia reciente. El 9 de noviembre de 1979, un problema informático llevó al NORAD (el Comando de Defensa Aeroespa- cial de América del Norte) a hacer un informe falso de un inminente ataque soviético a gran escala a Estados Unidos. Los EE.UU. hicieron los preparativos para una represalia de emergencia antes de que los datos de los sistemas de radar de alerta temprana mostraran que ningún ataque había sido lanzado (McLean y Stewart, 1979). El 26 de septiembre de 1983, el defectuoso sistema de alerta nuclear soviético Oko informó de un inminente ataque con misiles desde Estados Unidos. El informe fue identificado correctamente como una falsa alarma por el oficial de guardia en el centro de mando, Stanislav Petrov: una decisión que se ha acreditado que previno una guerra termonuclear (Lebedev, 2004). Parece que esa guerra probablemente no habría llegado a causar la extinción humana, aunque se hubiera luchado con los arsenales combinados de todas las

potencias nucleares en el apogeo de la Guerra Fría, a pesar de que habría arruinado la civilización y habría causado inimaginable muerte y sufrimiento (Gaddis, 1982; Parrington, 1997). Pero reservas más grandes podrían ser acumuladas en futuras carreras armamentís- ticas, o podrían inventarse armas incluso más letales, o nuestros modelos de impacto de un Armagedón nuclear (en particular, de la gravedad del invierno nuclear consecuente) podrían estar equivocados.

13. Este enfoque podría adaptarse a la categoría de método de control basado en la especificación directa de normas directas.
14. La situación es esencialmente la misma si el criterio de solución especifica una medida de bondad en lugar de un punto agudo de corte que cuente como solución.
15. Un defensor del enfoque del oráculo podría insistir en que hay al menos una posibilidad de que el usuario pudiera detectar el fallo en la solución ofrecida —reconociendo que no coincide con la intención del usuario, incluso aunque satisficiera los criterios de éxito especificados formalmente. La posibilidad de controlar el error en esta etapa dependerá de varios factores, incluyendo la forma humanamente comprensible de datos de salida del oráculo y de lo caritativo que sea al seleccionar cómo presentar al usuario el potencial resultado y sus factores decisivos.

    Alternativamente, en lugar de confiar en que el propio oráculo ofrezca estas funcionalidades, se podría tratar de construir una herramienta independiente que hiciera esto, una herramienta que podría inspeccionar los pronunciamientos del oráculo y mostrarnos de una manera útil lo que sucedería si actuamos sobre ellos. Pero hacer que esto de manera global requeriría otro oráculo superinteligente en cuyas adivinaciones entonces tendríamos que confiar; por lo que el problema de la fiabilidad no se habría resuelto, solamente se habría desplazado. Se podría tratar de obtener un incremento de la seguridad mediante el uso de múltiples oráculos que llevaran a cabo una revisión por pares, pero esto no protege del caso en que todos los oráculos fallan de la misma manera, como puede ocurrir si, por ejemplo, a todos ellos se les hubiera dado la misma especificación formal de lo que constituye una solución satisfactoria.
16. Bird y Layzell (2002) y Thompson (1997); También Yaeger (1994, 13-14).
17. Williams (1966).
18. Leigh (2010).
19. Este ejemplo está tomado de Yudkowsky (2011).
20. Wade (1976). Los experimentos en computadoras también se han realizado mediante una evolución simulada diseñada para parecerse en ciertos aspectos a la evolución biológica —de nuevo con resultados a veces extraños (véase, por ejemplo, Yaeger [1994]).
21. Con una cantidad suficientemente grande —finita pero físicamente improbable— de potencia de cálculo, probablemente *sería* posible lograr la superinteligencia general con los algoritmos disponibles en la actualidad. (Cf., por ejemplo, el sistema AIXItl;. Hutter [2001]). Pero incluso si la ley de Moore continuara por otros cien años, no sería suficiente para alcanzar los niveles de potencia de cálculo necesarios para lograrlo.

CAPÍTULO 11: ESCENARIOS MULTIPOLARES

1. No porque esto sea necesariamente el escenario más probable o más deseable, sino porque es el más fácil de analizar con el kit de herramientas de la teoría económica estándar, y por lo tanto un punto de partida para nuestra discusión.
2. American Horse Council (2005). Ver también Salem y Rowan (2001).
3. Acemoglu (2003); Mankiw (2009); Zuleta (2008).
4. Fredriksen (2012, 8); Salverda et al. (2009, 133).
5. También es esencial que al menos algo del capital sea invertido en activos que se eleven con la marea general. Una cartera de activos diversificada, tales como las acciones de un fondo indexado, aumentaría las posibilidades de no errar completamente.
6. Muchos de los sistemas de bienestar europeos están *sin fondos,* lo que significa que las pensiones se pagan con las cotizaciones y los impuestos en

curso de los trabajadores actuales y no de un depósito de ahorro. Tales esquemas no cumplirían automáticamente el requerimiento —en caso de un desempleo masivo repentino, los ingresos a partir de los cuales se pagan los beneficios podrían secarse. Sin embargo, los gobiernos pueden optar por compensar el déficit con otras fuentes.

7. American Horse Council (2005).
8. Proporcionar a siete mil millones de personas una pensión anual de $ 90,000 costaría $ 630 miles de millones al año, que es diez veces el PIB mundial actual. Durante los últimos cien años, el PIB mundial ha aumentado aproximadamente diecinueve veces de alrededor de $ 2 billones de dólares en 1900 hasta 37 miles de millones en 2000 (en dólares de 1990), según Maddison (2007). Así que si las tasas de crecimiento que hemos presenciado en los últimos cien años continuaran durante los siguientes doscientos años, mientras que la población se mantuviera constante, proporcionar a todo el mundo una pensión anual de $ 90.000 costaría alrededor del 3% del PIB mundial. Una explosión de inteligencia podría hacer que esta cantidad de crecimiento ocurriera en un lapso de tiempo mucho más corto. Véase también Hanson (1998a, 1998b, 2008).
9. Y quizás tanto como un millón de veces en los últimos 70.000 años, si hubo un estrechamiento pobla- cional severo alrededor de ese tiempo, como se ha especulado. Véase Kremer (1993) y Huff et al. (2010) para más información.
10. Cochran y Harpending (2009). Véase también Clark (2007) y, para una crítica, Alien (2008).
11. Kremer (1993).
12. Basten et al. (2013). Los escenarios en los que hay un aumento continuo también son posibles. En general, la incertidumbre de tales proyecciones aumenta en gran medida cuando vamos más allá de una o dos generaciones en el futuro.
13. Tomada globalmente, la tasa global de fecundidad de reemplazo fue de 2,33 hijos por mujer en 2003. Este número proviene del hecho de que se necesitan dos hijos por mujer para sustituir a los padres, además de un "tercer niño" para compensar (1) la probabilidad más alta de que nazcan niños varones, y (2) la mortalidad temprana antes del final de su vida fértil. Para los países desarrollados, el número es menor, en torno a 2,1, a causa de las tasas de mortalidad más bajas. (Ver Espenshade et al. [2003, Introducción, Tabla 1, 580]). La población en la mayoría de los países desarrollados se reduciría si no fuera por la inmigración. Algunos ejemplos notables de países con las tasas de fertilidad con sub-reemplazo son: Singapur en 0,79 (el más bajo en el mundo), Japón en 1,39, República Popular de China en 1,55, la Unión Europea en 1,58, Rusia en 1,61, Brasil en 1,81, Irán en 1,86, Vietnam a 1,87, y el Reino Unido en el 1,90. Incluso la población de Estados Unidos probablemente disminuya ligeramente con una tasa de fecundidad de 2,05. (Véase la CIA [2013]).
14. La plenitud de los tiempos puede ocurrir miles de millones de años a partir de ahora.
15. Carl Shulman señala que si los humanos biológicos cuentan con vivir su vida útil natural junto a la economía digital, tienen que asumir no sólo que el orden político en el ámbito digital protegería de los intereses humanos, sino que permanecería así durante períodos muy largos de tiempo (Shulman, 2012). Por ejemplo, si los acontecimientos en el ámbito digital se desplegaran miles de veces más rápido que en el exterior, entonces un ser humano biológico tendría que confiar en que el cuerpo político digital se mantuviera estable durante 50.000 años de cambio interno y agitación. Sin embargo, si el mundo político digital fuera algo parecido al nuestro, habría un gran número de revoluciones, guerras y conmociones catastróficas durante esos milenios que probablemente incomodarían a los humanos biológicos en el exterior. Incluso un riesgo de 0,01% por año de una guerra termonuclear global o de un cataclismo semejante implicaría una cierta

pérdida casi segura para los humanos biológicos que vivieran sus vidas a cámara lenta en tiempo sideral. Para superar este problema, sería necesario un orden más estable en el ámbito digital: tal vez una Unidad que mejorara gradualmente su propia estabilidad.

16. Uno podría pensar que incluso si las máquinas fueran mucho más eficientes que los seres humanos, todavía habría un cierto nivel de salarios en los que sería rentable emplear a un trabajador humano; por ejemplo, a 1 centavo por hora. Si ésta fuera la única fuente de ingresos para los seres humanos, nuestra especie se extinguiría ya que los seres humanos no pueden sobrevivir con 1 centavo por hora. Pero los humanos también obtienen ingresos de capital. Ahora bien, si asumimos que la población creciera hasta que el ingreso total estuviera a nivel de subsistencia, podría pensarse que en este estado los humanos estarían trabajando duro. Por ejemplo, supongamos que los ingresos de subsistencia fueran de 1$/día. Entonces, podría parecer que la población crecería hasta que la renta per cápita creciera hasta que la renta por persona proporcionada fuera de sólo unos 90 centavos de dólar por día de ingresos, lo que la gente tendría que complementar con diez horas de trabajo duro para compensar los 10 centavos restantes. Sin embargo, esto no tiene que ser así, porque los ingresos de subsistencia dependen de la cantidad de trabajo que se lleva a cabo: los seres humanos más trabajadores queman más calorías. Supongamos que cada hora de trabajo aumentan los costes de los alimentos en 2 centavos. Tenemos entonces un modelo en el que los seres humanos están ociosos en equilibrio.
17. Podría pensarse que un grupo político tan debilitado sería incapaz de votar y defender sus derechos. Pero los habitantes podrían dar poder a IAs fiduciarias para que gestionaran sus asuntos y representaran sus intereses políticos. (Esta parte de la discusión en esta sección se basa en la suposición de que se respetaran los derechos de propiedad).
18. No está claro cuál es el mejor término. “Matar” puede sugerir la brutalidad más activa de la que se pretende. “Fin” puede ser demasiado eufemístico. Una complicación es que hay dos eventos potencialmente separados: dejar de ejecutar activamente un proceso, y borrar la plantilla de información. Una muerte humana implica normalmente ambos eventos, pero para una emulación pueden estar separados. Que un programa deje *temporalmente* de ejecutarse puede ser igual de irrelevante que dormir para un ser humano; pero cesar *definitivamente* su funcionamiento puede ser el equivalente a entrar en un estado de coma permanente. Aún más complicaciones surgen del hecho de que las emulaciones pueden ser copiadas y funcionar a diferentes velocidades: posibilidades sin análogos directos en la experiencia humana. (Cfr Bostrom [2006b];. Bostrom y Yudkowsky [de próxima publicación]).
19. Habrá un equilibrio entre el poder de computación paralelo total y la velocidad de cálculo, ya que las más altas velocidades de computación serán alcanzables sólo a costa de una reducción en la eficiencia energética. Esto será especialmente cierto tras entrar en la era de la computación reversible.
20. Una emulación podría ser probada llevándola hacia alguna tentación. Al poner a prueba repetidamente cómo una emulación reacciona a partir de un cierto estado preprogramado a diversas secuencias de estímulos, se podría obtener una alta confianza en la fiabilidad de esa emulación. Pero cuanto más se le permita al estado mental desarrollarse lejos de su punto de partida validado, menos seguros podríamos estar de que seguiría siendo fiable. (En particular, ya que una emulación inteligente podría suponer que a veces estaría en una simulación, uno tendría que tener cuidado con la extrapolación de su comportamiento a situaciones en las que su hipótesis de simulación pesara menos en su toma de decisiones).
21. Algunas emulaciones podrían identificarse con su clan —es decir la totalidad de sus copias y variaciones derivadas de la misma plantilla— más

bien que con cualquier copia en particular. Tal emulación podría no considerar su propia eliminación como un evento de muerte, si supiera que otros miembros del clan sobrevivirían. Las emulaciones pueden saber que serán revertidas a un estado almacenado en particular al final del día y perder los recuerdos de ese día, pero no estaría desilusionado por esto como el asistente a la fiesta que sabe que va a despertar a la mañana siguiente sin ningún recuerdo de la noche anterior: entendería esto como amnesia retrógrada, no como la muerte.

22. Una evaluación ética podría tener en cuenta muchos otros factores. Incluso si todos los trabajadores estuvieran constantemente bien satisfechos con su condición, el resultado aún podría ser profundamente objetable en términos morales por otros motivos —aunque la base de estos otros motivos es una cuestión de conflicto entre teorías morales enfrentadas. Pero cualquier evaluación plausible consideraría el bienestar subjetivo como un factor importante. Véase también Bostrom y Yudkowsky (de próxima publicación).
23. World Values Survey (2008).
24. Helliwell y Sachs (2012).
25. Cf. Bostrom (2004). Ver también Chislenko (1996) y Moravec (1988).
26. Es difícil decir si las estructuras de procesamiento de información que surgirían en este tipo de escenario serían conscientes (en el sentido de tener experiencia fenoménica cualitativa). La razón de que esto sea difícil depende en parte de nuestra ignorancia empírica sobre las entidades cognitivas que surgirían y en parte de nuestra ignorancia filosófica sobre qué tipos de estructura tienen conciencia. Se podría tratar de replantear la cuestión, y en lugar de preguntarse si las entidades futuras serían conscientes, uno podría preguntarse si las entidades futuras tendrían estatus moral; o uno podría preguntarse si serían tales que tendríamos preferencias respecto de su "bienestar". Pero estas preguntas pueden no ser más fáciles de contestar que la pregunta acerca de la conciencia; de hecho, pueden llegar a requerir una respuesta a la pregunta conciencia por cuanto el estado moral o nuestras preferencias dependerán de que la entidad en cuestión pueda experimentar subjetivamente su condición.
27. Para un argumento que defiende que tanto en la historia geológica como en la humana se manifiesta esta tendencia hacia una mayor complejidad, véase Wright (2001). Para un argumento de oposición (criticado en el capítulo 9 del libro de Wright), véase Gould (1990). Veáse también Pinker (2011) para un argumento de que estamos asistiendo a una sólida tendencia a largo plazo de reducción de la violencia y la brutalidad.
28. Para más información sobre la teoría de la selección observacional, véase Bostrom (2002a).
29. Bostrom (2008a). Un examen mucho más cuidadoso de los detalles de nuestra historia evolutiva sería necesario para eludir el efecto de selección. Véase, por ejemplo, Carter (1983, 1993); Hanson (1998d); Cirkovic et al. (2010).
30. Kansa(2003).
31. Por ejemplo, Zahavi y Zahavi (1997).
32. Véase Miller (2000).
33. Kansa (2003). Para una aproximación provocativa, véase Frank (1999).
34. No es obvia la mejor manera de medir el grado de integración política mundial. Una perspectiva sería que, mientras que una tribu de cazadores-recolectores podría haber integrado un centenar de personas en una entidad de toma de decisiones, las entidades políticas más grandes de hoy en día contienen más de mil millones de personas. Esto equivaldría a una diferencia de siete órdenes de magnitud, con una sola magnitud adicional restante antes de que toda la población mundial estuviera contenida dentro de una sola entidad política. Sin embargo, en el momento en el que la tribu era la mayor escala de integración, la población mundial era mucho más pequeña. La tribu podría haber contenido tanto como una milésima parte de las personas que vivían entonces. Esto haría que el aumento en la escala

de integración política no fuera más que de dos órdenes de magnitud. Fijarse en la fracción de la población mundial que está políticamente integrada, en lugar de usar números absolutos, parece apropiado en el contexto actual (en particular debido a que la transición a la inteligencia artificial podría causar una explosión demográfica, de emulaciones u otras mentes digitales). Pero también ha habido avances en las instituciones y redes de colaboración a nivel mundial fuera de las estructuras formales del Estado, que también deben tenerse en cuenta.

35. Una de las razones para suponer que la primera revolución en inteligencia artificial será rápida —la posible existencia de un excedente de hardware— no se aplica en este caso. Sin embargo, podría haber otras fuentes de ganancia rápida, como un avance espectacular en el software asociado a la transición de una emulación puramente sintética a una inteligencia artificial.
36. Shulman (2010b).
37. Cómo se equilibrarían los pros y contras podría depender de qué tipo de trabajo estuviera tratando de hacer el superorganismo, y cómo de capaz fuera la plantilla de emulación más avanzada disponible. Parte de la razón de que se necesiten muchos tipos diferentes de seres humanos en las grandes organizaciones de hoy es que los seres humanos que destacan en múltiples ámbitos son raros.
38. Por supuesto, es muy fácil hacer varias copias de un agente de software. Pero téngase en cuenta que, en general, copiar no nos asegura suficientemente que las copias tengan los mismos objetivos finales. Para que dos agentes lleguen a tener los mismos objetivos finales (en un sentido relevante de "mismo"), los objetivos deben coincidir en sus elementos básicos. Si Bob es egoísta, una copia de Bob será probablemente egoísta. Sin embargo, sus objetivos no coinciden: Bob se preocupa por Bob mientras que la copia de Bob se preocupa por la copia de Bob.
39. Shulman (2010b, 6).
40. Esto podría ser más factible para los humanos biológicos y las emulaciones de cerebro completo que para inteligencias artificiales arbitrarias, que podrían construirse de manera que tuvieran compartimentos ocultos o dinámicas funcionales que podrían ser muy difíciles de descubrir. Por otro lado, las IAs construidas específicamente para ser transparentes deberían permitir una inspección y verificación más minuciosa de la que es posible con arquitecturas semejantes a los cerebros. Las presiones sociales pueden alentar a las IAs a exponer su código fuente, y a modificarse a sí mismas para ser transparentes —especialmente si ser transparente es una condición previa para ser de confianza y, por tanto, para que se les diera la oportunidad de participar en transacciones beneficiosas. Cf. Hall (2007).
41. Algunos otros temas que parecen relativamente menores, especialmente en casos en los que lo que está en juego es enorme (como lo son los fallos fundamentales de coordinación global), incluyen el coste de la búsqueda de encontrar políticas que podrían ser de mutuo interés, y la posibilidad de que algunos agentes pudieran tener una preferencia básica por la "autonomía" en una forma que se vería reducida mediante la firma de tratados integrales globales que contaran con mecanismos de monitorización e imposición.
42. Una IA quizá podría lograr esto mediante la modificación de sí misma apropiadamente y luego dando a los observadores un acceso de sólo lectura a su código fuente. Una inteligencia artificial con una arquitectura más opaca (como una emulación) tal vez podría lograrlo aplicándose a sí misma de manera pública algún método de selección de la motivación. Alternativamente, una agencia coercitiva externa, como un superorganismo policial, tal vez podría utilizarse no sólo para hacer cumplir la aplicación de un tratado celebrado entre los diferentes partidos, sino también internamente por un solo partido que se comprometa a un determinado curso de acción.

43. La selección evolutiva podría haber favorecido a los que ignoraban las amenazas e incluso a personajes visiblemente tan altamente inadaptables que prefirieran luchar hasta la muerte antes que sufrir la más mínima incomodidad. Tal disposición podría dar beneficios de señalización valiosos al portador. (Cualquiera de estas recompensas instrumentales por tener una necesidad disposicional no desempeñan, por supuesto, ningún papel en la motivación consciente del agente: éste podría valorar la justicia o el honor como fines en sí mismos).
44. Un veredicto definitivo sobre estas cuestiones, sin embargo, tendrá que esperar a un análisis posterior. Hay varias otras potenciales complicaciones que no podemos explorar aquí.

CAPÍTULO 12: ADQUIRIENDO VALORES

1. Se podrían introducir varias complicaciones y modulaciones de esta idea básica. Hablamos de una variación en el capítulo 8 —la de un agente satisfaciente, en lugar de maximizador— y en el capítulo siguiente tocamos brevemente la cuestión de las teorías de la decisión alternativas. Sin embargo, estas cuestiones no son esenciales para la idea central de este apartado, por lo que mantendremos las cosas simples, centrándonos aquí en el caso de un agente maximizador de la utilidad esperada.
2. Suponiendo que la IA tuviera una función de utilidad no trivial. Sería muy fácil construir un agente que eligiera siempre una acción que maximizara la utilidad esperada si su función de utilidad fuera, por ejemplo, la función constante U(w) = 0. Cada acción maximizaría igualmente bien la utilidad esperada para dicha función de utilidad.
3. También porque nos hemos olvidado de la ruidosa confusión de nuestra primera infancia, un momento en que aún no podíamos ver bien porque nuestro cerebro aún no había aprendido a interpretar su entrada visual.
4. Véase también Yudkowsky (2011) y la revisión de la sección 5 de Muehlhauser y Helm (2012).
5. Es quizás imaginable que los avances en ingeniería de software eventualmente pudieran superar estas dificultades. Usando herramientas modernas, un programador puede producir software que habría estado fuera del alcance de un equipo importante de desarrolladores que se hubieran visto obligados a escribir directamente en código computacional. Los programadores de IA de hoy ganan expresividad gracias a la amplia disponibilidad de aprendizaje de alta calidad artificial y de las bibliotecas de cálculo científico, que permiten a cualquiera piratear, por ejemplo, una aplicación de reconocimiento de rostros por cámara web mediante el encadenamiento de varias bibliotecas que nunca podría haber escrito por su cuenta. La acumulación de software reutilizable, producido por los especialistas, pero utilizables por los no especialistas, dará a los programadores futuros una ventaja en expresividad. Por ejemplo, un programador robótico futuro podría tener fácil acceso a las bibliotecas de impronta facial estándar, a colecciones típicas de oficina-edificio-objeto, a bibliotecas muy especializadas y a muchas otras funcionalidades que actualmente no están disponibles.
6. Dawkins (1995, 132). La afirmación aquí no es necesariamente que la cantidad de sufrimiento en el mundo natural *sea mayor* que la cantidad de bienestar positivo.
7. Los tamaños de población necesarios podrían ser mucho mayores o mucho menores que los que existían en nuestro propio linaje. Ver Shulman y Bostrom (2012).
8. Si fuera fácil conseguir un resultado equivalente sin perjudicar a un gran número de inocentes, parece moralmente mejor hacerlo. No obstante, si las personas digitales se crearan y se les infligiera un daño injusto, puede ser posible compensarlas por su sufrimiento archivándolas y volviendo a ejecutarlas más tarde (cuando el futuro de la humanidad estuviera asegurado) bajo condiciones más favorables. Ese resarcimiento podría compararse en algunos aspectos con las concepciones religiosas de una vida futura en el contexto de los intentos teológicos por abordar el problema de la evidencia del mal.
9. Una de las figuras principales de este campo, Richard Sutton, define el aprendizaje por refuerzo no en términos de un método de aprendizaje, sino en términos de un problema de aprendizaje: cualquier método que se adapte bien a la solución de ese

problema se considera un método de aprendizaje por refuerzo (Sutton y Barto, 1998, 4). La discusión actual, en cambio, se refiere a métodos en los que el agente puede ser concebido como teniendo el objetivo final de maximizar (alguna noción de) recompensa acumulada. Puesto que un agente con una especie muy diferente de objetivo final podría ser hábil imitando a un agente buscador de recompensas en una amplia gama de situaciones, y por lo tanto podría ser muy adecuado para la solución de los problemas de aprendizaje por refuerzo, podría haber métodos que contaran como "métodos de aprendizaje por refuerzo" según la definición de Sutton, pero que no darían lugar a un síndrome pinchazo cerebral. Las observaciones en el texto, sin embargo, se aplican a la mayoría de los métodos empleados en la comunidad de aprendizaje por refuerzo.

10. Incluso si, de alguna manera, un mecanismo similar a un ser humano pudiera establecerse dentro del intelecto artificial, los objetivos finales de origen humano adquiridos por este intelecto no tienen por qué asemejarse a los de un ser humano normal, a menos que el entorno de la crianza de este bebé digital también sea muy parecido al de un niño común: algo que sería difícil de organizar. E incluso con un ambiente de crianza similar al humano, no se puede garantizar un resultado satisfactorio, ya que incluso una diferencia sutil en disposiciones innatas podría dar lugar a reacciones muy diferentes en la vida real. Sin embargo, puede que sea posible crear un mecanismo más confiable de acumulación de valores para las mentes similares a las humanas en el futuro (tal vez usando nuevos fármacos o implantes cerebrales, o sus equivalentes digitales).
11. Uno podría preguntarse por qué los seres humanos no estamos tratando de desactivar el mecanismo que nos lleva a adquirir nuevos valores finales. Varios factores pueden estar en juego. En primer lugar, el sistema de la motivación humana está mal descrito como un algoritmo que maximiza la utilidad fríamente calculada. En segundo lugar, nosotros podríamos no tener ningún medio conveniente para alterar las formas en que adquirimos valores. En tercer lugar, es posible que tengamos razones instrumentales (que se entienden, por ejemplo, a partir de las necesidades de señalización social) para adquirir nuevos valores en ocasiones —los valores instrumentales podrían no ser tan útiles si nuestras mentes fueran parcialmente transparentes a otras personas, o si la complejidad cognitiva de simular tener un conjunto diferente de valores finales de lo que realmente se tiene es demasiado exigente. En cuarto lugar, hay casos en los que sí resistimos activamente tendencias que producen cambios en nuestros valores finales, por ejemplo cuando tratamos de resistir la influencia corruptora de las malas compañías. En quinto lugar, existe la interesante posibilidad de que diéramos algún valor final a ser el tipo de agente que pudiera adquirir nuevos valores finales por métodos humanos normales.
12. O uno podría tratar de diseñar un sistema de motivación en que la IA fuera indiferente a tal sustitución, véase Armstrong (2010).
13. Aquí vamos a recurrir a algunas aclaraciones hechas por Daniel Dewey (2011). Otras ideas de fondo que contribuyen a este marco han sido desarrolladas por Marcus Hutter (2005) y Shane Legg (2008), Eliezer Yudkowsky (2001), Nick Hay (2005), Moshe Looks, y Peter de Blanc.
14. Para evitar complicaciones innecesarias, limitamos nuestra atención a agentes deterministas que no descartan futuras recompensas.
15. Matemáticamente, el comportamiento de un agente puede formalizarse como una *función de agente,* que relaciona cada posible historia de interacción a una acción. A excepción de los agentes muy simples, no es factible representar la función de un agente explícitamente como una tabla de búsqueda. En su lugar, se da al agente alguna forma de calcular la acción a realizar. Puesto que hay muchas formas de calcular la misma función de agente, esto conduce a una individuación más fina del agente como *programa agente.* Un programa agente es un programa específico o un algoritmo que calcula la acción de cualquier historial de interacción dado. Aunque a menudo es matemáticamente conveniente y útil pensar en un programa agente que interactúa con algún entorno especificado formalmente, es importante recordar que se trata de una idealización. Los agentes reales están físicamente condicionados. Esto no sólo significa que el agente interactúa con el medio ambiente a través de sus

sensores y efectores, sino también que el "cerebro" del agente o controlador *es en sí mismo parte de la realidad física.* Sus operaciones pueden, por lo tanto, en principio, ser afectadas por interferencias físicas externas (y no sólo mediante la recepción de las percepciones de sus sensores). En algún momento, por lo tanto, se hace necesario ver al agente como un agente de aplicación. Un agente de implementación es una estructura física que, en ausencia de interferencia en su entorno, implementa una función de agente. Esta definición sigue a Dewey [2011]).

16. Dewey propone la siguiente noción optimizante para un agente de aprendizaje de valores:

$$y = \arg\max_{y_t \hat{} P, (yx_s,, \backslash yx_t y)} 2\, U(yxJP_2(U\backslash yxJ$$

Aquí, Pj y P_2 son dos funciones de probabilidad. El segundo sumatorio se extiende sobre alguna clase adecuada de funciones de utilidad sobre posibles historias de interacción. En la versión presentada en el texto, hemos hecho explícitas algunas dependencias, así como nos hemos aprovechado de las simplificaciones de notación disponibles.

17. Cabe señalar que el conjunto de funciones de utilidad U debe ser tal que las utilidades puedan ser comparadas y se promediadas. En general, esto es problemático, y no siempre es obvio cómo representar diferentes teorías morales de lo bueno en términos de funciones de utilidad cardinales. Véase, por ejemplo, MacAskill, 2010).

18. O en términos más generales, ya que *V* no podría ser tal que implicara directamente para cualquier par dado de un mundo posible y para una función de utilidad (w, U) si la proposición *V* (U) fuera cierta en w, lo que habría que hacer es dar a la IA una representación adecuada de la distribución de probabilidad condicional P(V *(U)\w).*

19. Consideremos primero Y, la clase de acciones que son posibles para un agente. Un problema aquí es decidir qué debería exactamente contar como una acción: ¿sólo las órdenes motoras básicas (por ejemplo, "envía un impulso eléctrico a lo largo del canal de salida # 00101100"), o las acciones de nivel superior (por ejemplo, "mantener la cámara centrada en la cara")? Ya que estamos tratando de desarrollar una noción de optimización en lugar de un plan de aplicación práctica, podemos entender que el dominio sean las órdenes motoras básicas (y como el conjunto de posibles órdenes motoras podría cambiar con el tiempo, es posible que necesitemos referir Y al tiempo). Sin embargo, con el fin de avanzar hacia la aplicación, es de suponer que será necesario introducir algún tipo de proceso de planificación jerárquico, y uno podría entonces considerar cómo aplicar la fórmula a alguna clase de acciones de nivel superior. Otra cuestión es cómo analizar acciones internas (como escribir cadenas a la memoria de trabajo). Dado que las acciones internas pueden tener consecuencias importantes, sería ideal desear que Y incluyera este tipo de acciones internas básicas, así como las órdenes motoras. Pero hay límites a lo lejos que se puede ir en esta dirección: el cálculo de la utilidad esperada de cualquier acción en Y requiere múltiples operaciones computacionales, y si cada una de esas operaciones también se considerara como una acción en Y que debía ser evaluada de acuerdo a la IA-VL, nos enfrentaríamos a una regresión infinita que haría imposible empezar. Para evitar el regreso al infinito, uno debe restringir cualquier intento explícito de estimar la utilidad esperada a un número limitado de posibilidades de acción importantes. El sistema entonces necesitaría algún proceso heurístico que identificara las posibilidades de acción importantes que deberían reexaminarse. (Con el tiempo el sistema también podría llegar a tomar decisiones explícitas respecto a algunas acciones posibles para hacer modificaciones a este proceso heurístico, acciones que podrían haber sido marcadas para que fueran atendidas explícitamente por este proceso de auto-referencia, de modo que a la larga el sistema podría llegar a ser cada vez más eficaz en la aproximación al ideal identificado por la IA-VL).

Consideremos a continuación W, que es una clase de mundo posible. Una dificultad aquí es especificar W de modo que sea suficientemente inclusivo. No incluir algunos w relevantes en W podría hacer que la IA fuera incapaz de representar una situación que se produjera en la realidad, lo que resulta en que la IA tomara malas decisiones. Supongamos, por ejemplo, que utilizamos una teoría ontológica para determinar la composición de W. Por ejemplo, incluimos en W

todos los mundos posibles, que consisten en un cierto tipo de colector de espacio-tiempo poblado por partículas elementales que se encuentran en el modelo estándar de la física de partículas. Esto podría distorsionar la epistemología de la IA si el modelo estándar fuera incompleto o incorrecto. Podríamos tratar de usar una clase más grande de W para cubrir más posibilidades; pero incluso si se pudiera asegurar que cada universo físico posible fuera incluido, todavía podríamos preocuparnos de que alguna otra posibilidad quedara fuera. Por ejemplo, ¿qué pasa con la posibilidad de mundos posibles dualistas en los que los hechos acerca de la conciencia no se convirtieran en hechos de la física? ¿Qué pasa con los hechos indicativos? ¿Con los hechos normativos? ¿Con los datos de las matemáticas superiores? ¿Qué pasa con otras clases de hechos que los seres humanos falibles podrían haber pasado por alto, pero que podrían llegar a ser importante para hacer que las cosas fueran tan bien como sea posible? Algunas personas tienen convicciones fuertes de que alguna teoría ontológica en particular es correcta. (Entre las personas que escriben sobre el futuro de la IA, la creencia en una ontología materialista, en la que la afirmación de que lo mental se traslada a lo físico, a menudo se da por sentada). Sin embargo, un momento de reflexión sobre la historia de las ideas debe ayudarnos a darnos cuenta de que hay una posibilidad significativa de que nuestra ontología favorita esté equivocada. Si los científicos del siglo XIX hubieran intentado una definición de inspiración física de W, probablemente habrían olvidado incluir la posibilidad de un espacio-tiempo no euclidiano o un Everettian ("una multitud de mundos") o la teoría cuántica o un multiverso cosmológico o la hipótesis de simulación —posibilidades que ahora parecen tener una probabilidad sustancial de obtenerse en el mundo real. Es plausible que haya otras posibilidades a las que la actual generación es igualmente ajena. (Por otro lado, si W es demasiado grande, surgen dificultades técnicas relacionadas con la necesidad de asignar medidas a conjuntos transfinitos). Lo ideal sería que pudiéramos disponer de alguna manera las cosas de modo que la IA pudiera utilizar algún tipo de ontología indefinida, una que la propia IA posteriormente podría extender usando los mismos principios que usaríamos a la hora de decidir si se debe reconocer un nuevo tipo de posibilidad metafísica.

Consideremos *P(w|Ey)*. La especificación deesta probabilidad condicional noes estrictamente parte del problema de introducción de valores. Con el fin de ser inteligente, la IA ya debería tener alguna manera de derivar probabilidades razonablemente exactas de muchas posibilidades fácticas relevantes. Un sistema que fuera demasiado lejos en este sentido no representaría el tipo de peligro que aquí nos ocupa. Sin embargo, puede haber un riesgo de que la IA terminara con una epistemología lo suficientemente buena como para hacer la IA instrumentalmente efectiva aunque no lo suficientemente buena como para que pudiera pensar correctamente acerca de algunas posibilidades de gran importancia normativa. (El problema de especificar *P(w\Ey)* está relacionado así con el problema de especificar W). La especificación de *P(w\Ey)* también requiere enfrentarse a otras cuestiones, como la forma de representar la incertidumbre sobre imposibilidades lógicas.

Las cuestiones antes citadas —cómo definir una clase de acciones posibles, una clase de mundos posibles y una distribución de probabilidad que conecte evidencia a las clases de mundos posible— son bastante genéricas: problemas similares surgen para una amplia gama de agentes formalmente especificados. Queda por examinar una serie de cuestiones más propias del enfoque de aprendizaje de valores; a saber, cómo definir U, *V(U),* y *P (V(U)\w)*.

U es una clase de funciones de utilidad. Hay una conexión entre U y W enla medida en que cada función de utilidad *U(w)* en U debería idealmente asignar utilidades a cada mundo posible *w* en W. Pero U también debe ser amplia en el sentido de que contenga funciones de utilidad suficientemente numerosas y diversas como para que nosotros tuviéramos una confianza justificada de que al menos uno de ellos fuera a hacer un buen trabajo representando los valores previstos.

La razón para escribir *P (V(U)\w)* en lugar de simplemente *P(U\w)* es para hacer hincapié en el hecho de que las probabilidades se asignan a proposiciones. Una función de utilidad, per se, no es una propuesta, pero podemos transformar una

función de utilidad en una propuesta al hacer alguna reclamación al respecto. Por ejemplo, podemos afirmar que una función de utilidad particular, U(.), describe las preferencias de una persona en particular, o que representa las prescripciones que implican cierta teoría ética, o que es la función de utilidad que el director hubiera deseado haber implementado si hubiera pensado en ello a fondo. El "criterio de valor" V(.) puede, por lo tanto, interpretarse como una función que toma como argumento una función de utilidad *U* y da como valor una propuesta de *U* que satisface el criterio *V*. Una vez que hemos definido una propuesta *V*(U), se espera que pudiéramos obtener la probabilidad condicional *P (V(U)\w)* de cualquier fuente que usáramos para obtener las otras distribuciones de probabilidad de la IA. (Si estamos seguros de que todos los hechos normativamente relevantes se fueran a tener en cuenta al individualizar los mundos posibles W, entonces *P (V(U)\w)* debería ser igual a cero o a uno en cada mundo posible). La pregunta sigue siendo cómo definir V. Esto se discute más adelante en el texto.

20. Éstos no son los únicos retos para el enfoque de aprendizaje de valores. Otra cuestión, por ejemplo, es cómo conseguir que la IA tenga unas creencias iniciales suficientemente sensatas, al menos por el tiempo que se vuelve lo suficientemente fuerte como para subvertir los intentos de los programadores para corregirla.
21. Yudkowsky (2001).
22. El término proviene del fútbol americano, donde un "Hail Mary" es un pase muy largo hacia adelante hecho a la desesperada, por lo general, cuando el tiempo está casi cumplido, buscando la remota posibilidad de que un jugador de su propio equipo atrape la pelota cerca de la zona de anotación y marque un *touchdown*.
23. El enfoque Hail Mary se basa en la idea de que una superinteligencia podría articular sus preferencias con mayor exactitud de lo que los seres humanos podemos articular las nuestras. Por ejemplo, una superinteligencia podría especificar su preferencia en código. Así que si nuestra IA representa a otras superinteligencias como procesos computacionales que se perciben su entorno, entonces nuestra IA debe ser capaz de razonar acerca de cómo esas superinteligencias extrañas responderían a algún estímulo hipotético, como que apareciera una "ventana" en su campo visual con el código fuente de nuestra propia IA pidiendo que especificaran sus instrucciones para nosotros en algún formato convenientemente preespecificado. Nuestra IA podría entonces leer estas instrucciones imaginarias (a partir de su propio modelo de escenario hipotético en el que estas superinteligencias extrañas están representadas), y nosotros hubiéramos construido nuestra IA para que estuviera motivada a seguir esas instrucciones.
24. Una alternativa sería la creación de un detector que buscara (en modelo mundial de nuestra AI) unas (representaciones de) estructuras físicas creadas por alguna civilización superinteligente. Entonces podríamos pasar por alto la etapa de identificación de funciones de preferencia de superinteligencias hipotéticas, y dar a nuestra propia IA el valor final de tratar de copiar las estructuras físicas que creyera que civilizaciones superinteligentes tenderían a producir.

    También hay desafíos técnicos con esta versión, sin embargo. Por ejemplo, puesto que nuestra propia IA, incluso después de haber alcanzado la superinteligencia, podría no ser capaz de conocer con gran precisión qué estructuras físicas construyen otras superinteligencias, nuestra IA podría necesitar recurrir a aproximaciones de esas estructuras. Para ello, nuestra IA probablemente necesitaría una métrica de similitud para juzgar cómo de cerca un artefacto físico se aproxima a otro. Pero las métricas de similitud basadas en medidas físicas brutas pueden ser inadecuadas — que sean apropiadas, por ejemplo, para juzgar si un cerebro se parece más a un queso camembert o a un ordenador ejecutando una emulación.

    Un enfoque más factible podría ser la búsqueda de "balizas": mensajes sobre funciones de utilidad codificados en un formato sencillo y adecuado. Construiríamos nuestra IA para que quisiera seguir cualquier tipo de mensaje acerca de funciones de utilidad que hipotetice pueda existir ahí fuera en el universo; y esperaríamos que las IAs extraterrestres amigables crearan una variedad de faros que ellos (con su superinteligencia) calcularan que las civilizaciones simples como la nuestra buscaríamos mediante la IA que fuéramos capaces de construir.
25. Si cada civilización tratara de resolver el problema de introducción de valores a

través de un Hail Mary, el pase sería un fracaso. Algunos tendrían que ponerse a hacerlo de la manera difícil.

26. Christiano (2012).
27. La IA que construyamos tampoco necesita ser capaz de encontrar el modelo. Al igual que nosotros, podría razonar sobre lo que una definición implícita tan compleja implicaría (tal vez mirando a su entorno y siguiendo la misma clase de razonamiento que estábamos siguiendo).
28. Cf. Capítulos 9 y 11.
29. Por ejemplo, el MDMA puede aumentar temporalmente la empatía; la oxitocina puede aumentar temporalmente la confianza (Vollenweider y col., 1998; Bartz et al., 2011). Sin embargo, los efectos parecen bastante variables y dependientes del contexto.
30. Los agentes mejorados podrían ser exterminados o colocados en animación suspendida (pausa), restablecidos a un estado anterior, o desposeídos de poder y frenados respecto de mejoras adicionales, hasta que el sistema global hubiera llegado a un estado más maduro y seguro donde estos elementos corruptos anteriores no representaran una amenaza para todo el sistema.
31. El problema también podría ser menos evidente en una sociedad futura de humanos biológicos, una que tuviera acceso a una vigilancia avanzada o a técnicas biomédicas para la manipulación psicológica, o que fuera lo suficientemente rica como para permitirse una proporción muy alta de profesionales de seguridad que vigilaran a la ciudadanía regular (y unos a otros).
32. Cf. Armstrong (2007) y Shulman (2010b).
33. Una pregunta abierta es hasta qué punto se necesita un supervisor de nivel n controlar no sólo a los supervisados de su nivel (n - 1), sino también a los de su nivel (n - 2), con el fin de saber qué nivel (n - 1) de agentes está haciendo su trabajo correctamente. Y para saber que el nivel (n - 1) de agentes ha logrado vigilar con éxito el nivel (n - 2) de agentes, ¿es necesario que el agente de nivel n monitorice también a los agentes de nivel (n - 3)?
34. Este enfoque se mueve entre la selección de motivación y el control de capacidad. Técnicamente, la parte de la disposición consiste en que los seres humanos que controlan un conjunto de supervisores de software cuente como control de capacidad, mientras que la parte de la disposición que consiste en capas de agentes de software dentro del sistema de control de otras capas es la selección de motivación (en la medida en que es un acuerdo que da forma a las tendencias motivacionales del sistema).
35. De hecho, muchos otros costes merecen consideración, pero no se pueden tratar aquí. Por ejemplo, fueran quienes fueran los agentes encargados de gobernar tal jerarquía, podrían volverse corruptos o envilecerse por su poder.
36. Para que esta garantía sea efectiva, se debería implementar de buena fe. Esto descartaría ciertos tipos de manipulación de las facultades emocionales y la toma de decisiones de la emulación que de otro modo podrían ser utilizadas (por ejemplo) para instalar un miedo a ser detenido o para evitar que la emulación evaluara racionalmente sus opciones.
37. Véase, por ejemplo, Brinton (1965); Goldstone (1980, 2001). (El progreso de las ciencias sociales sobre estas cuestiones podría ser un buen regalo para los déspotas del mundo, que podrían utilizar modelos predictivos más precisos de malestar social para optimizar sus estrategias de control de la población y para cortar suavemente las insurgencias que brotaran con menos fuerza letal).
38. Cf. Bostrom (2011a, 2009b).
39. En el caso de un sistema completamente artificial, podría ser posible obtener algunas de las ventajas de una estructura institucional sin crear realmente subagentes distintos. Un sistema podría incorporar múltiples perspectivas en su proceso de decisión sin dotar a cada uno de esos puntos de vista con su propia panoplia de facultades cognitivas necesarias para la agencia independiente. Podría ser difícil, sin embargo, aplicar plenamente el principio de “observar las consecuencias del comportamiento de un cambio propuesto, y volver a una versión anterior si las consecuencias parecieran indeseables desde el punto de vista anterior” descrito en el texto en un sistema que no estuviera compuesto de subagentes.

1. Un sondeo reciente entre filósofos profesionales sacó conclusiones sobre el porcentaje en que los en- cuestados "aceptaban o se inclinaban hacia" varias posiciones. En la ética normativa, los resultados fueron *deontología,* 25,9%; *consecuencialismo,* 23,6%; ética de la virtud, 18,2%. En metaética, los resultados fueron *realismo moral,* 56,4%; *antirrealismo moral,* 27,7%. En juicio moral: *cognitivismo,* 65,7%; *no cognitivismo,* 17,0% (Bourget y Ch Almers, 2009).
2. Pinker (2011).
3. Para una discusión sobre este asunto, véase Shulman et al. (2009).
4. Moore (2011).
5. Bostrom (2006b).
6. Bostrom (2009b).
7. Bostrom (2011a).
8. Más precisamente, deberíamos aplazar su opinión, excepto en aquellos temas en los que tenemos buenas razones para suponer que nuestras creencias son más precisas. Por ejemplo, podríamos saber más acerca de lo que estamos pensando en un momento particular de lo que sabría la superinteligencia si no fuera capaz de escanear el cerebro. Sin embargo, podríamos omitir esta calificación si asumiéramos que la superinteligencia tendría acceso a nuestras opiniones; podríamos entonces también dejar a la superinteligencia la tarea de juzgar cuando nuestras opiniones son dignas de confianza. (Pueden persistir algunos casos especiales, los que impliquen información indicativa, que necesite ser manejada por separado —por ejemplo, haciendo que la superinteligencia nos explicara qué sería racional creer desde nuestra perspectiva). Para introducirse en la floreciente literatura filosófica sobre el testimonio y autoridad epistémica, véase, por ejemplo, Elga (2007).
9. Yudkowsky (2004).Véase también Mijic (2010).
10. Por ejemplo, David Lewis propuso una *teoría disposicional del valor,* que sostiene, más o menos, que algo X es un valor para A si y sólo si A querría X si A fuera perfectamente racional y conociera perfectamente X (Smith et al., 1989). Ideas similares se habían presentado anteriormente; véase, por ejemplo, Sen y Williams (1982), Railton (1986), y Sidgwick y Jones (2010). A lo largo de las líneas algo similares, una forma común de justificación filosófica, *el método del equilibrio reflexivo,* propone un proceso de adaptación mutua iterativa entre nuestras intuiciones sobre casos particulares, las normas generales que creemos que gobiernan estos casos, y los principios según los cuales pensamos que estos elementos deberían ser revisados, para lograr un sistema más coherente; véase, por ejemplo, Rawls (1971) y Goodman (1954).
11. Es de suponer que la intención aquí es que cuando la IA actúe para prevenir este tipo de desastres, debe hacerlo *de la manera más suave posible*, es decir, de una manera tal que se evite el desastre, pero sin ejercer demasiada influencia sobre la humanidad en otros aspectos.
12. Yudkowsky (2004).
13. Rebecca Roache, comunicación personal.
14. Los tres principios son "Defender los seres humanos, el futuro de la humanidad y la naturaleza humana" (la integridad personal es aquí lo que nos gustaría ser, a diferencia de humanos, que es lo que somos); "La humanidad no debe pasar el resto de la eternidad deseando desesperadamente que los programadores hubieran hecho algo de manera diferente"; y "Ayudar a las personas".
15. Algunos grupos religiosos ponen un fuerte énfasis en la fe, en contraposición a la razón, a la que pueden llegar a considerar —incluso en su forma hipotéticamente más idealizada e incluso después de que se hubieran estudiado las escrituras, la revelación y la exégesis, con ardor y con mente abierta— como insuficiente para el logro de conocimientos espirituales esenciales. Aquellos que sostienen tales opiniones podrían no considerar la VCE como una guía óptima para la toma de decisiones (aunque podrían preferirla a varias otras guías imperfectas que podrían en realidad ser seguidas si se evitara el enfoque VCE).
16. Una IA que actuara como una fuerza latente de la naturaleza que regulara las interacciones humanas ha sido referida como un "Sysop," una especie de "sistema operativo" para la cuestión de la civilización humana. Véase Yudkowsky (2001).

17. "*Podría*", porque está *condicionado* con que la voluntad coherente extrapolada de la humanidad no deseara extender la consideración moral a estas entidades, por ser dudoso que esas entidades en realidad tuvieran estatus moral (a pesar de que parece muy plausible que lo tengan). "*Potencialmente*", porque incluso si un voto de bloqueo evita que la dinámica de la VCE dé protección directa a estos forasteros, todavía hay una posibilidad de que, dentro de lo que son las reglas básicas, se dejara de lado una vez que la dinámica inicial se hubiera parado, y que los individuos cuyos deseos fueran respetados y que quisieran un poco de bienestar para que los forasteros fueran protegidos y que pudieran negociar con éxito para lograr este resultado (a costa de renunciar a algunos de sus propios recursos). Que esto fuera posible podría depender, entre otras cosas, de que el resultado de la dinámica de la VCE fuera un conjunto de reglas de juego que hiciera posible llegar a una solución negociada de los problemas de este tipo (que podría requerir disposiciones para superar los problemas estratégicos de negociación).
18. Las personas que contribuyan positivamente a la realización de una superinteligencia segura y beneficiosa podrían merecer una recompensa especial por su trabajo, aunque algo menos que un poder casi exclusivo para determinar la disposición de los recursos cósmicos de la humanidad. Sin embargo, la noción de que todo el mundo obtuviera una parte igual en nuestra base de extrapolación es un buen punto de Schelling por lo que no debería ser desechado a la ligera. Hay, en todo caso, una manera indirecta para recompensar el trabajo bien hecho: a saber, que la propia VCE podría llegar a indicar que las buenas personas que defendieron a la humanidad en su nombre deberían ser reconocidas adecuadamente. Esto podría suceder sin que se les diera a estas personas ningún peso especial en la base de la extrapolación si —como es fácilmente imaginable— nuestra VCE aprobara (en el sentido de dar al menos un peso distinto a cero a) un principio justa recompensa.
19. Bostrom et al. (2013).
20. En la medida en que hay algunos significados compartidos (suficientemente precisos) que se expresa cuando hacemos afirmaciones morales, una superinteligencia debería ser capaz de averiguar cuál es ese significado. Y en la medida en que las afirmaciones morales son "susceptibles de ser verdaderas" (es decir, tienen un carácter proposicional subyacente que les permite ser verdaderas o falsas), la superinteligencia debería ser capaz de averiguar qué afirmaciones de la forma "Agente *X* debe ahora O" son verdaderas. Por lo menos, debería superarnos en esta tarea. Una IA que careciera inicialmente de tal capacidad para la cognición moral debería ser capaz de adquirirla si tuviera el superpoder de amplificación de la inteligencia. Una forma en la que la IA podría hacer esto es mediante la ingeniería inversa del pensamiento moral del cerebro humano para luego implementar un proceso similar pero ejecutándolo más rápido, alimentándolo con información fáctica más precisa y así sucesivamente.
21. Ya que no estamos seguros de la metaética, está la cuestión de qué podría hacer la IA si las condiciones previas para la RM no fueran alcanzadas. Una opción sería establecer que la IA se apagara a sí misma si asignara una probabilidad suficientemente alta a que el cognitivismo moral fuera falso o a que no hubiera verdades morales no-relativas adecuadas. Alternativamente, podríamos revertir la IA hasta cierto enfoque alternativo, como la VCE.

    Podríamos afinar la propuesta RM para clarificar lo que se va a hacer en varios casos ambiguos o degenerados. Por ejemplo, si la teoría del error es verdadera (y por lo tanto todas las afirmaciones morales positivas de la forma "Debo ahora O" son falsas), entonces la estrategia de retorno (por ejemplo, el apagado) se invocaría. También podríamos especificar lo que debería suceder si hubiera varias acciones posibles, cada una de los cuales fuera moralmente correcta. Por ejemplo, podríamos decir que en tales casos la IA debería realizar (una de) las medidas admisibles que la extrapolación colectiva de la humanidad habría favorecido. También podríamos estipular lo que debería suceder si la verdadera teoría moral no empleara términos como "moralmente correcto" en su vocabulario básico. Por ejemplo, una teoría consecuencialista podría sostener que algunas acciones son mejores que otras, pero que no existe un umbral particular, correspondiente a la noción de que una acción sea "moralmente correcta". Podríamos entonces decir que si esta teoría es correcta,

la RM debe realizar una de las mejores acciones moralmente factibles, si la hubiera; o, si hubiera un número infinito de acciones factibles tales que para cualquier acción viable hubiera una mejor, entonces tal vez la RM podría elegir cualquiera que fuera al menos astronómicamente mejor que la mejor acción que cualquier ser humano hubiera seleccionado en una situación similar, si tal acción fuera factible —o, en su defecto, una acción que fuera al menos tan buena como la mejor acción de un ser humano.

Hay que tener en cuenta un par de puntos generales a la hora de pensar en cómo podría refinarse la propuesta RM. En primer lugar, podríamos empezar de forma conservadora, utilizando la opción de reserva para cubrir casi todas las contingencias y sólo utilizar la opción de "corrección moral" en lAs que sentimos que entendemos completamente. En segundo lugar, podríamos añadir el modulador general a la propuesta de RM que debería "interpretar caritativamente, y revisarlo como lo hubiéramos revisado si hubiéramos pensado con más cuidado antes de que nos pusiéramos a escribirlo, etc.".

22. En estos términos, "conocimiento" puede parecer el término más fácilmente susceptible a un análisis formal (en términos de información teórica). Sin embargo, para representar lo que es para un humano saber algo, la IA puede necesitar un sofisticado conjunto de representaciones relacionadas con propiedades psicológicas complejas. Un ser humano no "sabe" toda la información que se almacena en alguna parte de su cerebro.
23. Un indicador de que los términos de la VCE serían (marginalmente) menos opacos es que contaría como progreso filosófico si pudiéramos analizar la rectitud moral en términos como los utilizados en la VCE. De hecho, una de las principales corrientes en metaética —la teoría de la observación ideal— propone hacer precisamente eso. Véase, por ejemplo, Smith et al. (1989).
24. Para ello es necesario enfrentar el problema de la incertidumbre normativa fundamental. Se puede demostrar que no siempre es adecuado actuar de acuerdo a la teoría moral que tiene la mayor probabilidad de ser cierta. También se podría demostrar que no siempre es adecuado realizar la acción que tiene mayor probabilidad de estar en lo correcto. Algunas formas de aumentar nuestras posibilidades frente a "grados de maldad" o frente a la gravedad de las cuestiones en juego parece ser necesaria. Para algunas ideas en este sentido, véase Bostrom (2009a).
25. Posiblemente podría incluso argumentarse que se trata de una condición de idoneidad para cualquier explicación sobre la noción de rectitud moral que explicara cómo Musculitos Joe es capaz de hacerse una idea de lo correcto e incorrecto.
26. No es obvio que lo moralmente correcto *para nosotros* sea construir una IA que implementara la RM, incluso si asumimos que *la propia IA* actuaría siempre moralmente. Tal vez sería objetablemente arrogante o prepotente por nuestra parte construir una IA de este tipo (sobre todo porque muchas personas podrían rechazar ese proyecto). Este problema puede afinarse parcialmente ajustando la propuesta RM. Supongamos que estipulamos que la IA debería actuar (hacer lo que sería moralmente correcto hacer) sólo si fuera moralmente correcto para sus creadores haber construido la IA en primer lugar; de lo contrario, debería apagarse. Es difícil ver cómo estaríamos cometiendo ningún mal moral grave al crear una IA de este tipo, ya que si hubiéramos hecho mal en crearla, la única consecuencia sería que se crearía una IA que inmediatamente se apagaría a sí misma, suponiendo que la IA no hubiera cometido ningún delito mental hasta ese momento. (Nosotros, sin embargo, podríamos haber actuado erróneamente —por ejemplo, al no haber aprovechado la oportunidad para construir algún otro tipo de IA en su lugar).

Una segunda cuestión es la supererrogación. Supongamos que hay muchas acciones que la IA pudiera tomar, cada una de las cuales sería moralmente correcta —en el sentido de ser *moralmente permisible*—, pero que algunas de ellas fueran moralmente mejores que las otras. Una opción es hacer que la IA se esfuerce en seleccionar la acción moralmente mejor en una situación semejante (o una de las mejores acciones, en caso de que haya varios que fueran igual de buenas). Otra opción es hacer que la IA seleccione de entre las acciones moralmente permisible

una que satisfaga al máximo algún otro desiderátum (no moral). Por ejemplo, la IA podría seleccionar, de entre las acciones que son moralmente permisibles, la acción que nuestra VCE preferiría tomar. Tal IA, sin hacer nada que no fuera moralmente permisible, podría proteger nuestros intereses más de lo que haría una IA que sólo hiciera lo que fuera moralmente mejor.

27. Cuando la IA evalúa la legitimidad moral de nuestro acto de creación de la IA, debería interpretar la permisibilidad en su sentido objetivo. En el sentido corriente de "moralmente permisible," un médico actúa de manera moralmente permisible cuando receta el medicamento que cree que va a curar a su paciente, incluso si el paciente, sin el conocimiento del médico, es alérgico a la droga y muere como resultado. Centrándose en la permisibilidad moral objetiva se aprovecha la posición epistémica presumiblemente superior de la IA.
28. Más directamente, depende de las *creencias* de la IA sobre qué teoría ética es verdadera (o, más precisamente, de su distribución de probabilidad entre las teorías éticas).
29. Puede ser difícil imaginar cómo de superlativamente maravillosas podrían ser estas vidas físicamente posibles. Véase Bostrom (2008c) para un intento poético de transmitir una idea de esto. Véase Bostrom (2008b) para el argumento de que algunas de estas posibilidades podrían ser buenas para nosotros, buenas para los seres humanos existentes.
30. Podría parecer engañoso o manipulador promover una de las propuestas si se piensa que alguna otra propuesta sería mejor. Pero uno podría promoverlo de manera que evitara la falta de sinceridad. Por ejemplo, se podría reconocer libremente la superioridad del ideal mientras se mantiene la promoción de la no ideal como el mejor compromiso posible.
31. O algún otro término positivamente evaluativo, como "bueno", "muy bueno" o "maravilloso".
32. Esto se hace eco de un principio en el diseño de software conocido como "Haz lo que quiero que hagas", o DWIM (Do What I Mean). Véase Teitelman (1966).
33. El contenido de los objetivos, la teoría de la decisión y la epistemología son tres aspectos que deben ser dilucidados; pero no tenemos la intención de afrontar la cuestión de si debe haber una descomposición ordenada de estos tres componentes por separado.
34. Un proyecto ético debería presumiblemente destinar como máximo una parte modesta de los eventuales beneficios que la superinteligencia produjera a recompensar especialmente a los que contribuyeron de manera moralmente permisible al éxito del proyecto. La asignación de una gran parte al entrelazado de incentivos sería indecoroso. Sería análogo a una organización benéfica que gastara el 90% de sus ingresos en bonos por desempeño para sus eventos de recaudación de fondos y en campañas de publicidad para aumentar las donaciones.
35. ¿Cómo podría recompensarse a los muertos? Podemos pensar en varias posibilidades. En el extremo inferior, podría haber plazas y monumentos conmemorativos, lo que sería una recompensa en la medida en que las personas desearan la fama póstuma. Los fallecidos también podrían tener otras preferencias sobre el futuro que podrían ser honradas, por ejemplo, en relación con las culturas, artes, edificios o entornos naturales. Por otra parte, la mayoría de las personas se preocupan por sus descendientes, y los privilegios especiales podían ser otorgados a los hijos y nietos de los contribuyentes. Más especulativamente, la superinteligencia podría ser capaz de crear simulaciones relativamente fieles de algunas personas pasadas — simulaciones que serían conscientes y que se asemejarían al original lo suficiente como para contar como una forma de supervivencia (de acuerdo con los criterios de algunas personas). Éste sería presumiblemente más fácil para las personas que hubieran entrado en suspensión criogénica; pero tal vez para una superinteligencia no sería imposible recrear algo muy similar a la persona original a partir de otros registros conservados como la correspondencia, publicaciones, materiales audiovisuales y documentos digitales, o los recuerdos personales de otros supervivientes. Una superinteligencia también podría pensar en algunas posibilidades que no se nos ocurrirían fácilmente a nosotros.
36. Sobre el atraco pascaliano, véase Bostrom (2009b). Para un análisis de las cuestiones

relacionadas con utilidades infinitas, véase Bostrom (2011a). Sobre incertidumbre normativa fundamental, véase, por ejemplo, Bostrom (2009a).

37. Por ejemplo, Price (1991); Joyce (1999); Drescher (2006); Yudkowsky (2010); Dai (2009).
38. Por ejemplo, Bostrom (2009a).
39. También es concebible que el uso de la normatividad indirecta para especificar el contenido objetivo de la IA mitigara los problemas que puedan surgir a partir de una teoría de la decisión especificada incorrectamente. Consideremos, por ejemplo, el enfoque de la VCE. Si se implementa bien, podría ser capaz de compensar, al menos, algunos errores en la especificación de la teoría de la decisión de la IA. La aplicación podría permitir a los valores que nuestra voluntad coherente extrapolada querría que la IA quisiera depender de la teoría de la decisión de la IA. Si nuestros seres idealizados supieran que estaban haciendo especificaciones de valor para una IA que estaba usando un tipo particular de teoría de la decisión, podrían ajustar sus especificaciones de valor de tal forma que la IA se comportase con benevolencia a pesar de su errónea teoría de la decisión —de manera muy similar a como se pueden anular los efectos distorsionadores de una lente colocando otra lente frente de ella que distorsiona de forma opuesta.
40. Algunos sistemas epistemológicos pueden no tener un fundamento claro en un sentido holístico. En ese caso, la herencia constitucional no es un conjunto claro y distinto de principios, sino más bien, por decirlo así, un punto de partida epistémico que encarna ciertas propensiones a responder a flujos entrantes de evidencia.
41. Véase, por ejemplo, el problema de distorsión discutido en Bostrom (2011a).
42. Por ejemplo, una cuestión controvertida en el razonamiento antrópico es si la llamada hipótesis de auto-referenciación debe ser aceptada. La hipótesis de auto-referenciación afirma, más o menos, que del hecho de que existes deberías inferir que las hipótesis según la cual existe un mayor número N de observadores debería recibir un impulso de probabilidad proporcional a N. Para un argumento en contra de este principio, véase el experimento mental del "Filósofo presuntuoso" en Bostrom (2002a). Para una defensa del principio, véase Olum (2002); y para una crítica de esa defensa, véase Bostrom y Cirkovic (2003). Las creencias sobre el supuesto de auto-referenciación pueden afectar a diversas hipótesis empíricas de importancia estratégica potencialmente crucial, por ejemplo, las consideraciones tales como el argumento del "día del juicio final" de Carter-Leslie, el argumento de simulación, y los argumentos de "gran filtro". Véase Bostrom (2002a, 2003a, 2008a); Carter (1983); Cirkovic et al. (2010); Hanson (1998d); Leslie (1996); Tegmark y Bostrom (2005). Algo similar podría hacerse en relación a otras tensas cuestiones de la teoría de la observación selectiva, como la cuestión de si la elección de la clase de referencia puede ser relativizada a sus momentos de observación, y si es así, cómo.
43. Véase, por ejemplo, Howson y Urbach (1993). También hay algunos resultados interesantes que reducen la gama de situaciones en las que dos agentes bayesianos pudieran estar racionalmente en desacuerdo cuando sus opiniones son de conocimiento común; véase AUmann (1976) y Hanson (2006).
44. Cf. el concepto de un "último juez" en Yudkowsky (2004).
45. Hay muchas cuestiones importantes pendientes en epistemología, algunas ya se mencionaron antes en el texto. El punto aquí es que es posible que no necesitemos acertar exactamente con todas las soluciones para lograr un resultado que fuera prácticamente indiscernible del mejor resultado. Un modelo mixto (que utilizara conjuntamente una amplia gama de diversos principios) podría funcionar.

CAPÍTULO 14: EL PANORAMA ESTRATÉGICO

1. Este principio se introdujo en Bostrom (2009b, 190), donde también se observó que no es tautológico. Para una analogía visual, imaginemos una caja con un volumen grande, pero finito, que representara el espacio de las capacidades básicas que podrían obtenerse a través de alguna posible tecnología. Imagínese que la arena que se vierte en esta caja representa el esfuerzo de investigación. Cómo se vierta la arena determina dónde se amontona en la caja. Pero si siguiéramos vertiendo la arena, el espacio entero finalmente se llenaría.

2. Bostrom (2002b).
3. Ésta no es la perspectiva desde la cual se ha contemplado la política sobre la ciencia tradicionalmente. Harvey Averch describe la política sobre ciencia y tecnología en Estados Unidos entre 1945 y 1984 como algo centrado en debates sobre el nivel óptimo de inversión pública en empresas de ciencia y tecnología, y en la medida en que el gobierno debe tratar de "escoger ganadores" con el fin de lograr el mayor aumento de la prosperidad económica de la nación y en la fuerza militar. En estos cálculos, el progreso tecnológico siempre se supone que es bueno. Pero Averch también describe el surgimiento de perspectivas críticas que cuestionan la premisa de que el "progreso es siempre bueno" (Averch, 1985). Véase también Graham (1997).
4. Bostrom (2002b).
5. Esto no es, por supuesto, tautológico de ninguna manera. Uno podría imaginar un argumento que defendiera un fin diferente de desarrollo. Se podría argumentar que sería mejor para la humanidad afrontar algún desafío menos difícil primero, por ejemplo el desarrollo de la nanotecnología, en base a que esto nos obligaría a desarrollar mejores instituciones, coordinarnos mejor internacionalmente y madurar nuestras ideas sobre estrategia mundial. Tal vez podríamos estar más dispuestos a afrontar un desafío que presentara una amenaza menos metafísicamente confusa que la superinteligencia artificial. La nanotecnología (o la biología sintética, o cualquiera que fuera el reto menor al que nos enfrentáramos primero) podría entonces servir como taburete que nos ayudaría a ascender al nivel de capacidad necesaria para hacer frente al desafío de mayor nivel de la superinteligencia.

   Tal argumento tendría que ser evaluado caso por caso. Por ejemplo, en el caso de la nanotecno- logía, tendrían que considerarse varias consecuencias posibles, como el aumento en el rendimiento del hardware de los sustratos computacionales nanofabricados; los efectos del capital físico barato de fabricación en el crecimiento económico; la proliferación de la tecnología de vigilancia sofisticada; la posibilidad de que un Unidad pudiera surgir a través de los efectos directos o indirectos de un gran avance nanotecnológico; y de la mayor viabilidad de las aproximaciones de la emulación de cerebro completo a la inteligencia artificial. Está más allá del alcance de nuestra investigación considerar todos estos temas (o los temas paralelos que puedan surgir para otras tecnologías que pudieran causar riesgos existenciales). Aquí sólo señalamos el argumento prima facie de favorecer una primera secuencia de desarrollo de la superinteligencia —mientras se hace hincapié en que hay complicaciones que podrían alterar esta evaluación preliminar en algunos casos.
6. Pinker (2011); Wright (2001).
7. Podría ser tentador suponer la hipótesis de que todo se ha acelerado hasta el punto de carecer de significado, sobre la base de que no parece (a primera vista) tener consecuencias observacionales; pero véase, por ejemplo, Shoemaker (1969).
8. El nivel de preparación no se mide por la cantidad de esfuerzo invertido en las actividades de preparación, sino por lo adecuadamente configuradas que estén realmente las condiciones y por lo bien posicionados que estén para elegir la acción apropiada los que vayan a tomar las decisiones clave.
9. El grado de confianza internacional durante el período previo a la explosión de inteligencia también podría ser un factor. Consideraremos esto en la sección de "Colaboración" más adelante en el capítulo.
10. Como anécdota, parece que aquellos que están seriamente interesados en la actualidad en el problema de control están situados de manera desproporcionada en un extremo superior de la distribución de inteligencia, aunque podría haber otras explicaciones para esta impresión. Si el campo se pone de moda, sin duda se verá inundado de mediocres y raritos.
11. Le debo este término a Carl Shulman.
12. ¿Cómo de similar a un cerebro tiene que ser una inteligencia artificial para contar como una emulación de cerebro completo en lugar de como una IA neuromórfica? El factor determinante podría ser que el sistema reprodujera cualquiera de los valores o la panoplia completa de tendencias cognitivas y eva- luativas de cualquier individuo en particular o de un ser humano genérico, porque esto probablemente marcaría la diferencia en el problema de control. Alcanzar estas propiedades podría

requerir un grado más alto de fidelidad en la emulación.

13. La magnitud del impulso, por supuesto, dependerá de lo grande que fuera el impulso, y también de la procedencia de los recursos de dicho impulso. Puede que no haya ningún aumento neto para la neu- rociencia si todos los recursos adicionales invertidos en la investigación sobre la emulación de cerebro completo se dedujeran de la investigación en neurociencia regular —a menos que un enfoque más agudo de investigación sobre la emulación acabara de descubrir una forma más efectiva de avanzar en la neurociencia que el camino actual de la investigación en neurociencias.
14. Véase Drexler (1986, 242). Drexler (comunicación privada) confirma que esta reconstrucción se corresponde con el razonamiento que él estaba tratando de presentar. Obviamente, una serie de premisas implícitas tendrían que ser añadidas si se quisiera emitir el argumento en forma de cadena de razonamiento deductivamente válida.
15. ¿No deberíamos quizás dar la bienvenida a pequeñas catástrofes si aumentaran nuestra vigilancia hasta el punto de que previnieran catástrofes de mediana escala que habrían sido necesarias para hacernos tomar las fuertes precauciones necesarias para evitar catástrofes existenciales? (Y, por supuesto, al igual que con los sistemas inmunes biológicos, también tenemos que estar preocupados con las sobre-reaccciones, de manera análoga a las alergias y a los trastornos autoinmunes).
16. Cf. Lenman (2000); Burch-Brown (2014).
17. Cf. Bostrom (2007).
18. Téngase en cuenta que este argumento se centra en el orden en lugar de en la cronología de los acontecimientos relevantes. Hacer que la superinteligencia llegue antes ayudaría a anticiparse a otros riesgos de transición existenciales sólo si la intervención cambiara la secuencia de los acontecimientos clave: por ejemplo, que la superinteligencia hiciera que diversos hitos en nanotecnología o en biología sintética se alcanzaran antes.
19. Si solucionar el problema de control fuera *extremadamente* difícil en comparación con la solución al problema de rendimiento de la inteligencia artificial, y si la capacidad del proyecto tuviera sólo una correlación débil con el tamaño del proyecto, entonces sería posible que fuera preferible que un pequeño proyecto llegara primero, es decir, si la variación en capacidad fuera mayor entre los proyectos más pequeños. En tal situación, incluso si los proyectos más pequeños fueran, en promedio, menos competentes que los proyectos más grandes, podría ser menos improbable que un pequeño proyecto pasara a tener el nivel monstruosamente alto de competencia necesario para resolver el problema de control.
20. Esto no niega que uno pueda imaginar herramientas que podrían promover la deliberación mundial y que se beneficiarían de, o incluso requerirían, seguir avanzando en el hardware —por ejemplo, la traducción de alta calidad, las búsquedas mejores, el acceso ubicuo a teléfonos inteligentes, los entornos de realidad virtual atractivos para las relaciones sociales, y así sucesivamente.
21. La inversión en tecnología de emulación podría acelerar el progreso hacia la emulación de cerebro completo no sólo directamente (a través de las prestaciones técnicas producidas), sino también indirectamente mediante la creación de un grupo de votantes que luchará por obtener más fondos y aumentar la visibilidad y la credibilidad de la visión de la emulación de cerebro completo (ECC).
22. ¿Cuánto valor esperado se perdería si el futuro estuviera conformado por los deseos de un ser humano al azar y no por (una superposición adecuada de) los deseos de toda la humanidad? Esto podría depender sensiblemente de la norma de evaluación que utilizáramos, y también de si los deseos en cuestión estuvieran idealizados o en bruto.
23. Por ejemplo, mientras que la mente humana se comunica lentamente a través del lenguaje, las IAs podrían diseñarse de manera que copias del mismo programa fueran capaces de transferir fácil y rápidamente habilidades e información entre ellas. Las mentes artificiales diseñadas *ab initio* podrían no necesitar los engorrosos sistemas heredados que ayudaron a nuestros antepasados a ocuparse de ciertos aspectos del entorno natural que son poco importantes en el ciberespacio. Las

mentes digitales también podrían diseñarse para aprovecharse del rápido procesamiento en serie disponible para cerebros biológicos, y de lo fácil que sería instalar nuevos módulos de funcionalidad altamente optimizada (por ejemplo, el procesamiento simbólico, el reconocimiento de patrones, los simuladores, la minería de datos y la planificación). La inteligencia artificial también podría tener ventajas no técnicas importantes, tales como la facilidad para ser patentada o su menor tendencia a enredarse en complejidades morales derivadas de la información humana.

24. Si p_1 y p_2 son las probabilidades de fracaso en cada paso, la probabilidad total de fallo es $p_1 + (1 - p_1)\, p_2$ ya que una puede fallar de manera terminal sólo una vez.
25. Es posible, por supuesto, que el proyecto aventajado no llegara a tener una ventaja tan grande y no fuera capaz de formar una Unidad. También es posible que surgiera una Unidad antes de la IA incluso sin la intervención de la ECC, en cuyo caso esta razón para favorecer el escenario en que aparezca primero la ECC se derrumba.
26. ¿Hay una manera para que un promotor de la ECC aumente la especificidad de su apoyo de tal manera que se acelere la ECC y reduzca al mínimo la propagación al desarrollo de la IA? Promover una tecnología de escaneo es probablemente una mejor opción que promover el modelado neurocomputacional. (Promover hardware probablemente no marcará mucho la diferencia en uno u otro sentido, dados los grandes intereses comerciales que de todos modos estarían incentivando el progreso en dicho campo).

    Promover la tecnología de escaneo puede aumentar la probabilidad de un resultado multipolar al hacer que el escaneo sea algo menos restringido, algo que aumentaría la probabilidad de que la primera población de emulaciones fuera reproducida a partir de muchas plantillas humanas diferentes en lugar de consistir en millones de copias hechas a partir de un pequeño número de plantillas. Los avances en la tecnología de escaneo también harían más probable que las restricciones estuvieran en el cálculo de hardware, lo que tendería a ralentizar el despegue.
27. Una IA neuromórfica también podría carecer de otras características de seguridad de la emulación de cerebro completo, como tener un perfil de fortalezas cognitivas y debilidades similares a las de un ser humano biológico (lo que nos permitiría usar nuestra experiencia con los seres humanos para formarnos expectativas respecto de las capacidades del sistema en diferentes etapas de su desarrollo).
28. Si el motivo de alguien para promover la ECC fuera hacer que la ECC sucediera antes de la IA, debería tenerse en cuenta que la aceleración de la ECC alteraría el orden de llegada sólo si el tiempo predeterminado de los dos caminos hacia la inteligencia artificial fuera parejo con una ligera ventaja para la IA. De lo contrario, la inversión en ECC simplemente haría que la ECC sucediera antes de lo que debería (reduciendo excedente de hardware y el tiempo de preparación), pero sin afectar a la secuencia de desarrollo; o bien la inversión en ECC tendría poco efecto (excepto provocar, quizás, que la IA llegara incluso antes, al estimular los progresos en IA neuromórfica).
29. Comentarios sobre esto en Hanson (2009).
30. Habría, por supuesto, *alguna* magnitud e inminencia de riesgo existencial para la que también fuera preferible, incluso desde la perspectiva de la persona afectada, posponer el riesgo —ya sea que las personas existentes puedan aprovechar un poco más la vida antes de que se cierre el telón, o para proporcionar más tiempo a preparar esfuerzos de mitigación que podrían reducir el peligro.
31. Supongamos que pudiéramos tomar alguna acción que acortara el plazo hasta la explosión de inteligencia en un año. Digamos que las personas que actualmente habitan la Tierra están muriendo a un ritmo del 1% por año, y que el riesgo por defecto de que la explosión de inteligencia produzca la extinción humana es del 20% (escogiendo arbitrariamente un número para ilustrar el argumento). En ese caso, acelerar la llegada de la explosión de inteligencia por un año podría compensar (desde el punto de vista de la persona afectada) aumentar el riesgo de 20% a 21%, es decir, aumentar en un 5% el nivel de riesgo. Sin embargo, la gran mayoría de personas vivas un año antes del inicio de la explosión de inteligencia que en ese momento tengan un interés en posponerlo para reducir en un punto

porcentual el riesgo de explosión (ya que la mayoría de los individuos creerían que su riesgo de morir el año que viene es mucho menor que el 1% —dado que la mayor mortalidad se produce en demografías relativamente restringidas, como las de los enfermos y la de los ancianos). Así, uno podría tener un modelo en el que cada año la población votara posponer la explosión de inteligencia por otro año, por lo que la explosión de inteligencia nunca sucedería, aunque todo el mundo que viviera estuviera de acuerdo en que sería mejor que la explosión de inteligencia sucediera en algún momento. En realidad, por supuesto, las fallas de coordinación, la predictibilidad limitada o las preferencias por cosas distintas a la supervivencia personal, probablemente impedirían una posposición indefinida.

Si se utiliza el factor de descuento económico estándar en lugar de la norma de la persona afectada, la magnitud del potencial de crecimiento disminuye, ya que el valor de que las personas existentes puedan llegar a disfrutar de vidas astronómicamente largas se rebaja de manera abrupta. Este efecto es especialmente fuerte si el factor de descuento se aplica al tiempo subjetivo de cada individuo y no a tiempo sideral. Si los beneficios futuros son descontados a una tasa del x% por año, y el nivel de riesgo existencial de fondo proveniente de otras fuentes es de un y % al año, entonces el punto óptimo para la explosión de inteligencia sería cuando la demora de la explosión por un año produjera menos de x + y puntos porcentuales de reducción del riesgo existencial asociado con una explosión de inteligencia.

32. Estoy en deuda con Carl Shulman y Stuart Armstrong por su ayuda con este modelo. Véase también Shulman (2010a, 3): "Chalmers (2010) informa de un consenso entre los cadetes y personal de la academia militar estadounidense de West Point de que el gobierno de Estados Unidos no iba a frenar la investigación en IA incluso frente a una potencial catástrofe, por temor a que las potencias rivales ganaran una ventaja decisiva".
33. Es decir, que la información sobre el modelo siempre es mala *ex ante.* Por supuesto, dependiendo de lo que la información realmente sea, en algunos casos podría llegar a ser bueno que la información se diera a conocer, sobre todo si la brecha entre el líder y el segundo fuera mucho mayor de lo que se hubiera creído razonablemente de antemano.
34. Podría incluso representar un riesgo existencial, especialmente si estuviera precedida por la introducción de nuevas tecnologías militares de destrucción o por la acumulación de armas sin precedentes.
35. Un proyecto podría tener a sus trabajadores distribuidos en un gran número de localidades colaborando a través de canales de comunicación cifrados. Pero esta táctica implica una disyuntiva de seguridad: si bien la dispersión geográfica podría ofrecer cierta protección contra los ataques militares, también podría obstaculizar la seguridad operativa, ya que es más difícil evitar que el personal deserte, que la información se filtre, o que fueran secuestrados por un poder rival si estuvieran distribuidos en muchos lugares.
36. Téngase en cuenta que un gran factor de rebaja temporal podría hacer que un proyecto se comportara en algunos aspectos como si estuviera en una carrera, aunque supiera que no tiene ningún competidor real. Un gran factor de descuento significaría que se preocuparía poco sobre el futuro lejano. Dependiendo de la situación, esto desalentaría a bluesky I + D, lo que tendería a retrasar la revolución de la inteligencia artificial (aunque tal vez haría que fuera más abrupta cuando ocurriera, a causa del excedente de hardware). Pero un gran factor de descuento —o un bajo nivel de preocupación respecto de las futuras generaciones— también podrían hacer creer que tomar riesgos existenciales importara menos. Esto alentaría apuestas que implicarían la posibilidad de una ganancia inmediata a expensas de un mayor riesgo de catástrofe existencial, desincentivando así la inversión en seguridad e incentivando un lanzamiento temprano —que imitaría los efectos de una carrera dinámica. A diferencia de la dinámica de carrera, sin embargo, un gran factor de descuento (o la indiferencia hacia las generaciones futuras) no tendrían ninguna tendencia particular a incitar al conflicto.

Reducir la dinámica de carrera es uno de los principales beneficios de la colaboración. Que la colaboración facilite el intercambio de ideas sobre cómo resolver el problema de control es también un beneficio, aunque esto está

contrarrestado en cierta medida por el hecho de que la colaboración también facilitaría el intercambio de ideas sobre cómo resolver el problema de la competencia. El efecto neto de esta facilitación en el intercambio de ideas puede ser un leve aumento en la inteligencia colectiva de la comunidad investigadora relevante.

37. Por otro lado, la supervisión pública por parte de un solo gobierno correría el riesgo de producir un resultado en el que una nación monopolizara las ganancias. Este resultado parece inferior a aquel en la que incontables altruistas garantizan que todo el mundo tenga mucho que ganar. Además, la supervisión por parte de un gobierno nacional no significaría necesariamente que todos los ciudadanos de ese país fueran a recibir una parte de la prestación: según el país de que se tratara, habría un mayor o menor riesgo de que todos los beneficios fueran capturados por una élite política o unos pocos agentes egoístas.
38. Un requisito sería que el uso del entrelazamiento de incentivos (como se discutió en el Capítulo 12) pudiera, en algunas circunstancias, animar a la gente a unirse a un proyecto como colaboradores activos y no como individualistas pasivos.
39. Los rendimientos decrecientes se situarían en una escala mucho menor. La mayoría de la gente preferiría tener una estrella que una probabilidad entre mil millones de tener una galaxia con mil millones de estrellas. De hecho, la mayoría de la gente preferiría tener una milmillonésima parte de los recursos de la Tierra que una posibilidad entre mil millones de ser dueño de todo el planeta.
40. Cf. Shulman (2010a).
41. Las teorías éticas agregativas se meten en problemas cuando tomamos en serio la idea de que el cosmos podría ser infinito; véase Bostrom (2011b). También puede haber problemas cuando tomamos en serio la idea de los valores ridículamente grandes pero finitos; véase Bostrom (2009b).
42. Si hiciéramos crecer un ordenador, finalmente nos enfrentaríamos a limitaciones relativistas que surgirían de las demoras de comunicación entre las diferentes partes del ordenador —las señales no se propagan más rápido que la luz. Si uno contrae el equipo, uno se encuentra con límites cuánticos a la miniaturización. Si se aumenta la densidad de la computadora, uno choca con el límite del agujero negro. Es cierto que no podemos estar completamente seguros de que no descubriremos una nueva física que nos ofrezca alguna manera de sobreponernos a estas limitaciones.
43. El número de copias de una persona podría escalar linealmente con recursos sin límite. Sin embargo, no está claro hasta qué punto el ser humano promedio valoraría tener varias copias de sí mismo. Incluso aquellas personas que prefirieran ser copiadas muchas veces podrían no tener una función de utilidad que fuera lineal con un número creciente de copias. El número de copias, al igual que los años de vida, pueden tener rendimientos decrecientes en función de la utilidad típica de cada persona.
44. Una Unidad es altamente colaborativa de manera interna en el caso de la toma de decisiones de alto nivel. Una Unidad podría tener grandes cantidades de no-colaboración y conflicto en los niveles inferiores, si el organismo de más alto nivel que constituye la Unidad eligiera mantener las cosas de esa manera.
45. Si cada equipo de desarrollo de IA rival está convencido de que los otros equipos están tan equivocados como para no tener ninguna posibilidad de producir una explosión de inteligencia, entonces una de las razones para colaborar —evitar la dinámica de carrera— es obviada: cada equipo debería elegir de forma independiente ir más lento confiando en la creencia de que carece de cualquier competencia seria.
46. Un estudiante de doctorado.
47. Esta formulación está pensada para ser leída como una receta para garantizar que se les dé la debida atención al bienestar de los animales no humanos y de otros seres sensibles (incluyendo las mentes digitales) que existan o puedan llegar a existir. No está destinado a ser leído como una licencia para que un desarrollador de IA sustituya sus propias intuiciones morales por las de una comunidad moral más amplia. El principio es coherente con el enfoque de la "voluntad coherente extrapolada" discutido en el capítulo 12, con una base de extrapolación que incluya a todos los seres humanos.

Una aclaración: La formulación no pretende excluir necesariamente la posibilidad de los derechos de propiedad post-transición de las superinteligencias artificiales o de sus algoritmos constituyentes y estructuras de datos. La formulación está destinada a ser agnóstica acerca de cómo se estructurarían los sistemas jurídicos o políticos para permitir las transacciones en una futura e hipotética sociedad posthumana. Lo que la formulación está destinada a afirmar es que la elección de un sistema de este tipo, en la medida en que su selección está determinada causalmente por cómo se desarrolló inicialmente la superinteligencia, debería hacerse sobre la base del criterio indicado; es decir, el sistema constitucional posterior a la transición debe ser elegido para el beneficio de toda la humanidad y al servicio de los ideales-éticos ampliamente compartidos en comparación con, por ejemplo, el beneficio exclusivo de los que desarrollaran la superinteligencia por primera vez.

48. Los refinamientos de la cláusula de bonanza son obviamente posibles. Por ejemplo, tal vez el umbral debe ser expresado en términos *per capita,* o tal vez al ganador se le debería permitir mantener algo más que una parte equitativa del excedente para incentivar con más fuerza la producción futura (una versión del principio maximin de Rawls podría ser atractivo en este caso). Otras mejoras podrían reorientar la cláusula en términos distintos a los económicos y replantearse en términos de "influencia en el futuro de la humanidad" o de "grado en que los intereses de las diferentes partes son tenidas en cuenta en función de utilidad de una futura Unidad" o algo así.

CAPÍTULO 15: LA HORA DE LA VERDAD

1. Algunas investigaciones valen la pena, no por lo que descubren, sino por otras razones, como por su capacidad para entretener, educar, otorgar reconocimiento, o para que los que participan en ella se sientan realizados.
2. No estoy sugiriendo que *nadie* deba trabajar en matemáticas puras o filosofía. Tampoco estoy sugiriendo que estos esfuerzos sean un desperdicio en comparación con todos las demás pérdidas de tiempo de la academia o de la sociedad en general. Probablemente es muy bueno que algunas personas puedan dedicarse a la vida intelectual y seguir su curiosidad intelectual dondequiera que les lleve, independiente de cualquier pensamiento utilitario. La sugerencia es que en el margen, algunas de las mejores mentes, al darse cuenta de que sus actuaciones cognitivas podrían llegar a quedar obsoletas en un futuro próximo, podrían querer desviar su atención de los problemas teóricos para marcar una diferencia en obtener la solución un poco antes.
3. Aunque uno debería ser prudente en los casos en que esta incertidumbre pudiera ser protectora —recordemos, por ejemplo, el modelo de carrera de riesgo del Cuadro 13, donde encontramos que la información estratégica adicional podría ser perjudicial. De manera más general, es necesario preocuparse acerca de los peligros de la información (véase Bostrom [2011b]). Es tentador decir que necesitamos más análisis de los riesgos de la información. Esto es probablemente cierto, aunque todavía podríamos preocuparnos de que tales análisis pudieran producir información peligrosa.
4. Cf. Bostrom (2007).
5. Agradezco a Carl Shulman que recalcara este punto.

1.

BIBLIOGRAFÍA

Acemoglu, Daron. 2003. "Labor- and Capital-Augmenting Technical Change." *Journal of the European Economic Association* 1 (1): 1-37.

Albertson, D. G., and Thomson, J. N. 1976. "The Pharynx of *Caenorhabditis Elegans." Philosophical Transac- tions of the Royal Society B: Biological Sciences* 275 (938): 299-325.

Allen, Robert C. 2008. "A Review of Gregory Clark's *A Farewell to Alms: A Brief Economic History of the World." Journal of Economic Literature* 46 (4): 946-73.

American Horse Council. 2005. "National Economic Impact of the US Horse Industry." Retrieved July 30, 2013. Available at http://www.horsecouncil.org/national-economic-impact-us-horse-industry.

Anand, Paul, Pattanaik, Prasanta, and Puppe, Clemens, eds. 2009. *The Oxford Handbook of Rational and Social Choice.* New York: Oxford University Press.

Andres, B., Koethe, U., Kroeger, T., Helmstaedter, M., Briggman, K. L., Denk, W., and Hamprecht, F. A. 2012. "3D Segmentation of SBFSEM Images of Neuropil by a Graphical Model over Supervoxel Boundaries." *Medical Image Analysis* 16 (4): 796-805.

Armstrong, Alex. 2012. "Computer Competes in Crossword Tournament." *I Programmer,* March 19.

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Unpublished manuscript, October 20. Retrieved December 31, 2012. Available at http://www. neweuropeancentury.org/GodAI.pdf.

Armstrong, Stuart. 2010. *Utility Indifference,* Technical Report 2010-1. Oxford: Future of Humanity Institute, University of Oxford.

Armstrong, Stuart. 2013. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* 12: 68-84.

Armstrong, Stuart, and Sandberg, Anders. 2013. "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox." *Acta Astronautica* 89: 1-13.

Armstrong, Stuart, and Sotala, Kaj. 2012. "How We're Predicting AI—or Failing To." In *Beyond AI: Artificial Dreams,* edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52-75. Pilsen: University of West Bohemia. Retrieved February 2, 2013.

Asimov, Isaac. 1942. "Runaround." *Astounding Science-Fiction,* March, 94-103.

Asimov, Isaac. 1985. *Robots and Empire.* New York: Doubleday.

Aumann, Robert J. 1976. "Agreeing to Disagree." *Annals of Statistics* 4 (6): 1236-9.

Averch, Harvey Allen. 1985. *A Strategic Analysis of Science and Technology Policy.* Baltimore: Johns Hopkins University Press.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Jacob, W., Lent, R., and Herculano-Houzel, S. 2009. "Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain." *Journal of Comparative Neurology* 513 (5): 532-41.

Baars, Bernard J. 1997. *In the Theater of Consciousness: The Workspace of the Mind.* New York: Oxford University Press.

Baratta, Joseph Preston. 2004. *The Politics of World Federation: United Nations, UNReform, Atomic Control.* Westport, CT: Praeger.

Barber, E. J. W. 1991. *Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Spe- cial Reference to the Aegean.* Princeton, NJ: Princeton University Press.

Bartels, J., Andreasen, D., Ehirim, P., Mao, H., Seibert, S., Wright, E. J., and Kennedy, P. 2008. "Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex." *Journal of Neuro- science Methods* 174 (2): 168-76.

Bartz, Jennifer A., Zaki, Jamil, Bolger, Niall, and Ochsner, Kevin N. 2011. "Social Effects of Oxytocin in Hu- mans: Context and Person Matter." *Trends in Cognitive Science* 15 (7): 301-9.

Basten, Stuart, Lutz, Wolfgang, and Scherbov, Sergei. 2013. "Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility." *Demographic Research* 28: 1145-66.

Baum, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge, MA: MIT Press.

Baum, Seth D., Goertzel, Ben, and Goertzel, Ted G. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1): 185-95.

Beal, J., and Winston, P. 2009. "Guest Editors' Introduction: The New Frontier of Human-Level Artificial Intelligence." *IEEEIntelligent Systems* 24 (4): 21-3.

Bell, C. Gordon, and Gemmell, Jim. 2009. *TotalRecall: How the E-Memory Revolution Will Change Everything.* New York: Dutton.

Benyamin, B., Pourcain, B. St., Davis, O. S., Davies, G., Hansell, M. K., Brion, M.-J. A., Kirkpatrick, R. M., et al. 2013. "Childhood Intelligence is Heritable, Highly Polygenic and Associated With FNBP1L." *Molecular Psychiatry* (January 23).

Berg, Joyce E., and Rietz, Thomas A. 2003. "Prediction Markets as Decision Support Systems." *Information Systems Frontiers* 5 (1): 79-93.

Berger, Theodore W., Chapin, J. K., Gerhardt, G. A., Soussou, W. V., Taylor, D. M., and Tresco, P. A., eds. 2008. *Brain-Computer Interfaces: An International Assessment of Research and Development Trends.* Springer.

Berger, T. W., Song, D., Chan, R. H., Marmarelis, V. Z., LaCoss, J., Wills, J., Hampson, R. E., Deadwyler, S. A., and Granacki, J. J. 2012. "A Hippocampal Cognitive Prosthesis: Multi-Input, Multi-Output Nonlinear Modeling and VLSI Implementation." *IEEE Transactions on Neural Systems and Rehabilitation Engineer- ing* 20 (2): 198-211.

Berliner, Hans J. 1980a. "Backgammon Computer-Program Beats World Champion." *Artificial Intelligence* 14 (2): 205-220.

Berliner, Hans J. 1980b. "Backgammon Program Beats World Champ." *SIGART Newsletter* 69: 6-9.

Bernardo, José M., and Smith, Adrian F. M. 1994. *Bayesian Theory*, 1st ed. Wiley Series in Probability & Sta- tistics. New York: Wiley.

Birbaumer, N., Murguialday, A. R., and Cohen, L. 2008. "Brain-Computer Interface in Paralysis." *Current Opinion in Neurology* 21 (6): 634-8.

Bird, Jon, and Layzell, Paul. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." In *Proceedings of the 2002 Congress on Evolutionary Computation,* 2: 1836-41.

Blair, Clay, Jr. 1957. "Passing of a Great Mind: John von Neumann, a Brilliant, Jovial Mathematician, was a Prodigious Servant of Science and His Country." *Life*, February 25, 89-104.

Bobrow, Daniel G. 1968. "Natural Language Input for a Computer Problem Solving System." In *Semantic Information Processing,* edited by Marvin Minsky, 146-227. Cambridge, MA: MIT Press.

Bostrom, Nick. 1997. "Predictions from Philosophy? How Philosophers Could Make Themselves Useful." Un- published manuscript. Last revised September 19, 1998.

Bostrom, Nick. 2002a. *Anthropic Bias: Observation Selection Effects in Science and Philosophy.* New York: Routledge.

Bostrom, Nick. 2002b. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9.

Bostrom, Nick. 2003a. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53

(211): 243-55.

Bostrom, Nick. 2003b. “Astronomical Waste: The Opportunity Cost of Delayed Technological Development.”
Utilitas 15 (3): 308-314.

Bostrom, Nick. 2003c. “Ethical Issues in Advanced Artificial Intelligence.” In *Cognitive, Emotive and Ethical Aspects of Decisión Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George E. Lasker, 2: 12-17. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cy- bernetics.

Bostrom, Nick. 2004. “The Future of Human Evolution.” In *Two Hundred Years After Kant, Fifty Years After Turing,* edited by Charles Tandy, 2: 339-371. Death and Anti-Death. Palo Alto, CA: Ria University Press. Bostrom, Nick. 2006a. “How Long Before Superintelligence?” *Linguistic and PhilosophicalInvestigations* 5(1): 11-30.

Bostrom, Nick. 2006b. “Quantity of Experience: Brain-Duplication and Degrees of Consciousness.” *Minds and Machines* 16 (2): 185-200.

Bostrom, Nick. 2006c. “What is a Singleton?” *Linguistic and Philosophical Investigations* 5 (2): 48-54. Bostrom, Nick. 2007. “Technological Revolutions: Ethics and Policy in the Dark.” In *Nanoscale: Issues and Perspectives for the Nano Century,* edited by Nigel M. de S. Cameron and M. Ellen Mitchell, 129-52. Ho- boken, NJ: Wiley.

Bostrom, Nick. 2008a. “Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing.” *MIT Technology Review,* May/June issue, 72-7.

Bostrom, Nick. 2008b. “Why I Want to Be a Posthuman When I Grow Up.” In *Medical Enhancement and Post- humanity,* edited by Bert Gordijn and Ruth Chadwick, 107-37. New York: Springer.

Bostrom, Nick. 2008c. “Letter from Utopia.” *Studies in Ethics, Law, and Technology* 2 (1): 1-7.

Bostrom, Nick. 2009a. “Moral Uncertainty - Towards a Solution?”. *Overcoming Bias* (blog), January 1. Bostrom, Nick. 2009b. “Pascal’s Mugging.” *Analysis* 69 (3): 443-5.

Bostrom, Nick. 2009c. “The Future of Humanity.” In *New Waves in Philosophy of Technology,* edited by Jan Kyrre Berg Olsen, Evan Selinger, and S0ren Riis, 186-215. New York: Palgrave Macmillan.

Bostrom, Nick. 2011a. “Information Hazards: A Typology of Potential Harms from Knowledge.” *Review of Contemporary Philosophy* 10: 44-79.

Bostrom, Nick. 2011b. “Infinite Ethics.” *Analysis and Metaphysics* 10: 9-59.

Bostrom, Nick. 2012. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.” In “Theory and Philosophy of AI,” edited by Vincent C. Müller, special issue, *Minds and Machines* 22 (2): 71-85.

Bostrom, Nick, and Cirkovic, Milan M. 2003. “The Doomsday Argument and the Self-Indication Assumption: Reply to Olum.” *Philosophical Quarterly* 53 (210): 83-91.

Bostrom, Nick, and Ord, Toby. 2006. “The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics.”
Ethics 116 (4): 656-79.

Bostrom, Nick, and Roache, Rebecca. 2011. “Smart Policy: Cognitive Enhancement and the Public Interest.” In *Enhancing Human Capacities*, edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane, 138-49. Malden, MA: Wiley-Blackwell.

Bostrom, Nick and Sandberg, Anders. 2009a. “Cognitive Enhancement: Methods, Ethics, Regulatory Chal- lenges.” *Science and EngineeringEthics* 15 (3): 311-41.

Bostrom, Nick and Sandberg, Anders. 2009b. “The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement.” In *Human Enhancement*, 1st ed., edited by Julian Savulescu and Nick Bostrom, 375-416. New York: Oxford University Press.

Bostrom, Nick, Sandberg, Anders, and Douglas, Tom. 2013. “The Unilateralist’s Curse: The Case for a Princi- ple of Conformity.” Working Paper. Retrieved February 28, 2013. Available at http://www.nickbostrom. com/papers/unilateralist.pdf.

Bostrom, Nick, and Yudkowsky, Eliezer. Forthcoming. “The Ethics of Artificial Intelligence.” In *Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.

Boswell, James. 1917. *Boswell’s Life of Johnson.* New York: Oxford University Press.

Bouchard, T. J. 2004. “Genetic Influence on Human Psychological Traits: A Survey.” *Current Directions in Psychological Science* 13 (4): 148-51.

Bourget, David, and Chalmers, David. 2009. “The PhilPapers Surveys.” November. Available at http:// philpapers.org/surveys/.

Bradbury, Robert J. 1999. “Matrioshka Brains.” Archived version. As revised August 16, 2004. Available at http://web.archive.org/web/20090615040912/http://www.aeiveos.com/~bradbury/MatrioshkaBrains/ MatrioshkaBrainsPaper.html.

Brinton, Crane. 1965. *The Anatomy of Revolution.* Revised ed. New York: Vintage Books.

Bryson, Arthur E., Jr., and Ho, Yu-Chi. 1969. *Applied Optimal Control: Optimization, Estimation, and Control.* Waltham, MA: Blaisdell.

Buehler, Martin, Iagnemma, Karl, and Singh, Sanjiv, eds. 2009. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*. Springer Tracts in Advanced Robotics 56. Berlin: Springer.

Burch-Brown, J. 2014. “Clues for Consequentialists.” *Utilitas* 26 (1): 105-19.

Burke, Colin. 2001. “Agnes Meyer Driscoll vs. the Enigma and the Bombe.” Unpublished manuscript. Re- trieved February 22, 2013. Available at http://userpages.umbc.edu/~burke/driscoll1-2011.pdf.

Canbáck, S., Samouel, P., and Price, D. 2006. “Do Diseconomies of Scale Impact Firm Size and Performance? A Theoretical and Empirical Overview.” *Journal of Managerial Economics* 4 (1): 27-70.

Carmena, J. M., Lebedev, M. A., Crist, R. E., O’Doherty, J. E., Santucci, D. M., Dimitrov, D. F., Patil, P. G., Henriquez, C. S., and Nicolelis, M. A. 2003. “Learning to Control a Brain-Machine Interface for Reaching and Grasping by Primates.” *Public Library of Science Biology* 1 (2): 193-208.

Carroll, Bradley W., and Ostlie, Dale A. 2007. An Introduction to Modern Astrophysics. 2nd ed. San Francisco: Pearson Addison Wesley.

Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Carter, Brandon. 1983.“The Anthropic Principle and its Implications for Biological Evolution.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 310 (1512): 347-63.

Carter, Brandon. 1993. “The Anthropic Selection Principle and the Ultra-Darwinian Synthesis.” In *The An- thropic Principle: Proceedings of the Second Venice Conference on Cosmology and Philosophy,* edited by F. Bertola and U. Curi, 33-66. Cambridge: Cambridge University Press.

CFTC & SEC (Commodity Futures Trading Commission and Securities & Exchange Commission). 2010. *Findings Regarding the Market Events of May 6,2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington, DC.

Chalmers, David John. 2010. “The Singularity: A Philosophical Analysis.” *Journal of Consciousness Studies* 17 (9-10): 7-65.

Chason, R. J., Csokmay, J., Segars, J. H., DeCherney, A. H., and Armant, D. R. 2011. “Environmental and Epi- genetic Effects Upon Preimplantation Embryo Metabolism and Development.” *Trends in Endocrinology and Metabolism* 22 (10): 412-20.

Chen, S., and Ravallion, M. 2010. “The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty.” *Quarterly Journal of Economics* 125 (4): 1577-1625.

Chislenko, Alexander. 1996. “Networking in the Mind Age: Some Thoughts on Evolution of Robotics and Distributed Systems.” Unpublished manuscript.

Chislenko, Alexander. 1997. “Technology as Extension of Human Functional Architecture.” *Extropy Online.*

Chorost, Michael. 2005. *Rebuilt: How Becoming Part Computer Made Me More Human.* Boston: Houghton Mifflin.

Christiano, Paul F. 2012. “‘Indirect Normativity’ Write-up.” *Ordinary Ideas* (blog), April 21.

CIA. 2013. “The World Factbook.” Central Intelligence Agency. Retrieved August 3. Available at https://www. cia.gov/library/publications/the-world-factbook/rankorder/2127rank.html? countryname=United%20States&countrycode=us&regionCode=noa&rank=121#us.

Cicero. 1923. “On Divination.” In *On Old Age, on Friendship, on Divination,* translated by W. A. Falconer. Loeb Classical Library. Cambridge, MA: Harvard University Press.

Cirasella, Jill, and Kopec, Danny. 2006. “The History of Computer Games.” Exhibit at Dartmouth Artificial Intelligence Conference: The Next Fifty Years (AI@50), Dartmouth

College, July 13-15.
Cirkovic, Milan M. 2004. “Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelli- gent Beings.” *Foundations of Physics* 34 (2): 239-61.
Cirkovic, Milan M., Sandberg, Anders, and Bostrom, Nick. 2010. “Anthropic Shadow: Observation Selection Effects and Human Extinction Risks.” *Risk Analysis* 30 (10): 1495-1506.
Clark, Andy, and Chalmers, David J. 1998. “The Extended Mind.” *Analysis* 58 (1): 7-19.
Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World.* 1st ed. Princeton, NJ: Prince- ton University Press.
Clavin, Whitney. 2012. “Study Shows Our Galaxy Has at Least 100 Billion Planets.” *Jet Propulsion Laboratory,* January 11.
CME Group. 2010. *What Happened on May 6th?* Chicago, May 10.
Coase, R. H. 1937. “The Nature of the Firm.” *Economica* 4 (16): 386-405.
Cochran, Gregory, and Harpending, Henry. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York: Basic Books.
Cochran, G., Hardy, J., and Harpending, H. 2006. “Natural History of Ashkenazi Intelligence.” *Journal of Biosocial Science* 38 (5): 659-93.
Cook, James Gordon. 1984. *Handbook of Textile Fibres: Natural Fibres*. Cambridge: Woodhead.
Cope, David. 1996. *Experiments in Musical Intelligence*. Computer Music and Digital Audio Series. Madison, WI: A-R Editions.
Cotman, Carl W., and Berchtold, Nicole C. 2002. “Exercise: A Behavioral Intervention to Enhance Brain Health and Plasticity.” *Trends in Neurosciences* 25 (6): 295-301.
Cowan, Nelson. 2001. “The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity.” *Behavioral and Brain Sciences* 24 (1): 87-114.
Crabtree, Steve. 1999. “New Poll Gauges Americans’ General Knowledge Levels.” *Gallup News,* July 6.
Cross, Stephen E., and Walker, Edward. 1994. “Dart: Applying Knowledge Based Planning and Scheduling to Crisis Action Planning.” In *Intelligent Scheduling,* edited by Monte Zweben and Mark Fox, 711-29. San Francisco, CA: Morgan Kaufmann.
Crow, James F. 2000. “The Origins, Patterns and Implications of Human Spontaneous Mutation.” *Nature Reviews Genetics* 1 (1): 40-7.
Cyranoski, David. 2013. “Stem Cells: Egg Engineers.” *Nature* 500 (7463): 392-4.
Dagnelie, Gislin. 2012. “Retinal Implants: Emergence of a Multidisciplinary Field.” *Current Opinion in Neu- rology* 25 (1): 67-75.
Dai, Wei. 2009. “Towards a New Decision Theory.” *Less Wrong* (blog), August 13.
Dalrymple, David. 2011. “Comment on Kaufman, J. ‘Whole Brain Emulation: Looking at Progress on *C. Ele- gans*.’ ” *Less Wrong* (blog), October 29.
Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X., et al. 2011. “Genome-Wide Associ- ation Studies Establish That Human Intelligence Is Highly Heritable and Polygenic.” *Molecular Psychiatry* 16 (10): 996-1005.
Davis, Oliver S. P., Butcher, Lee M., Docherty, Sophia J., Meaburn, Emma L., Curtis, Charles J. C., Simpson, Michael A., Schalkwyk, Leonard C., and Plomin, Robert. 2010. “A Three-Stage Genome-Wide Association Study of General Cognitive Ability: Hunting the Small Effects.” *Behavior Genetics* 40 (6): 759-767.
Dawkins, Richard. 1995. *River Out of Eden: A Darwinian View of Life.* Science Masters Series. New York: Basic Books.
De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents’ Value Systems.* Machine Intelligence Research Institute, San Francisco, CA, May 19.
De Long, J. Bradford. 1998. “Estimates of World GDP, One Million B.C.-Present.” Unpublished manuscript.
De Raedt, Luc, and Flach, Peter, eds. 2001. *Machine Learning: ECML 2001:12th European Conference on Machine Learning, Freiburg, Germany, September 5-7,2001. Proceedings.* Lecture Notes in Computer Science 2167. New York: Springer.
Dean, Cornelia. 2005. “Scientific Savvy? In U.S., Not Much.” *New York Times,* August 30.
Deary, Ian J. 2001. “Human Intelligence Differences: A Recent History.” *Trends in Cognitive Sciences* 5 (3): 127-30.

Deary, Ian J. 2012. “Intelligence.” *Annual Review of Psychology* 63: 453-82.

Deary, Ian J., Penke, L., and Johnson, W. 2010. “The Neuroscience of Human Intelligence Differences.” *Nature Reviews Neuroscience* 11 (3): 201-11.

Degnan, G. G., Wind, T. C., Jones, E. V., and Edlich, R. F. 2002. “Functional Electrical Stimulation in Tetraple- gic Patients to Restore Hand Function.” *Journal of Long-Term Effects of Medical Implants* 12 (3): 175-88.

Devlin, B., Daniels, M., and Roeder, K. 1997. “The Heritability of IQ.” *Nature* 388 (6641): 468-71.

Dewey, Daniel. 2011. “Learning What to Value.” In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings,* edited by Jürgen Schmid- huber, Kristinn R. Thórisson, and Moshe Looks, 309-14. Lecture Notes in Computer Science 6830. Berlin: Springer.

Dowe, D. L., and Hernández-Orallo, J. 2012. “IQ Tests Are Not for Machines, Yet.” *Intelligence* 40 (2): 77-81.

Drescher, Gary L. 2006. *Good and Real: DemystifyingParadoxes from Physics to Ethics.* Bradford Books. Cambridge, MA: MIT Press.

Drexler, K. Eric. 1986. *Engines of Creation*. Garden City, NY: Anchor.

Drexler, K. Eric. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation.* New York: Wiley.

Drexler, K. Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization.* New York: PublicAffairs.

Driscoll, Kevin. 2012. “Code Critique: ‘Altair Music of a Sort.’ ” Paper presented at Critical Code Studies Working Group Online Conference, 2012, February 6.

Dyson, Freeman J. 1960. “Search for Artificial Stellar Sources of Infrared Radiation.” *Science* 131 (3414): 16671668.

Dyson, Freeman J. 1979. *Disturbing the Universe.* 1st ed. Sloan Foundation Science Series. New York: Harper & Row.

Elga, Adam. 2004. “Defeating Dr. Evil with Self-Locating Belief.” *Philosophy and Phenomenological Research* 69 (2): 383-96.

Elga, Adam. 2007. “Reflection and Disagreement.” *Noüs* 41 (3): 478-502.

Eliasmith, Chris, Stewart, Terrence C., Choo, Xuan, Bekolay, Trevor, DeWolf, Travis, Tang, Yichuan, and Rasmussen, Daniel. 2012. “A Large-Scale Model of the Functioning Brain.” *Science* 338(6111): 1202-5.

Ellis, J. H. 1999. “The History of Non-Secret Encryption.” *Cryptologia* 23 (3): 267-73.

Elyasaf, Achiya, Hauptmann, Ami, and Sipper, Moche. 2011. “Ga-Freecell: Evolving Solvers for the Game of Freecell.” In *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference,* 19311938. GECCO’ 11. New York: ACM.

Eppig, C., Fincher, C. L., and Thornhill, R. 2010. “Parasite Prevalence and the Worldwide Distribution of Cognitive Ability.” *Proceedings of the Royal Society B: Biological Sciences* 277 (1701): 3801-8.

Espenshade, T. J., Guzman, J. C., and Westoff, C. F. 2003. “The Surprising Global Variation in Replacement Fertility.” *Population Research and Policy Review* 22 (5-6): 575-83.

Evans, Thomas G. 1964. “A Heuristic Program to Solve Geometric-Analogy Problems.” In *Proceedings of the April 21-23, 1964, Spring Joint Computer Conference,* 327-338. AFIPS ’64. New York: ACM.

Evans, Thomas G. 1968. “A Program for the Solution of a Class of Geometric-Analogy Intelligence-Test Ques- tions.” In *Semantic Information Processing*, edited by Marvin Minsky, 271-353. Cambridge, MA: MIT Press.

Faisal, A. A., Selen, L. P., and Wolpert, D. M. 2008. “Noise in the Nervous System.” *Nature Reviews Neuroscience* 9 (4): 292-303.

Faisal, A. A., White, J. A., and Laughlin, S. B. 2005. “Ion-Channel Noise Places Limits on the Miniaturization of the Brain’s Wiring.” *Current Biology* 15 (12): 1143-9.

Feldman, Jacob. 2000. “Minimization of Boolean Complexity in Human Concept Learning.” *Nature* 407 (6804): 630-3.

Feldman, J. A., and Ballard, Dana H. 1982. “Connectionist Models and Their Properties.” *Cognitive Science* 6 (3): 205-254.

Foley, J. A., Monfreda, C., Ramankutty, N., and Zaks, D. 2007. “Our Share of the Planetary Pie.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12585-6.

Forgas, Joseph P., Cooper, Joel, and Crano, William D., eds. 2010. *The Psychology of Attitudes and Attitude Change.* Sydney Symposium of Social Psychology. New York: Psychology Press.

Frank, Robert H. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess.* New York: Free Press.

Fredriksen, Kaja Bonesmo. 2012. *Less Income Inequality and More Growth - Are They Compatible?: Part 6. The Distribution of Wealth.* Technical report, OECD Economics Department Working Papers 929. OECD Publishing.

Freitas, Robert A., Jr. 1980. “A Self-Replicating Interstellar Probe.” *Journal of the British Interplanetary Society* 33: 251-64.

Freitas, Robert A., Jr. 2000. “Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations.” Foresight Institute. April. Retrieved July 28, 2013. Available at http://www. foresight.org/nano/Ecophagy.html.

Freitas, Robert A., Jr., and Merkle, Ralph C. 2004. *Kinematic Self-Replicating Machines.* Georgetown, TX: Landes Bioscience.

Gaddis, John Lewis. 1982. *Strategies of Containment: A Critical Appraisal of Postwar American National Secu- rity Policy.* New York: Oxford University Press.

Gammoned.net. 2012. “Snowie.” Archived version. Retrieved June 30. Available at http://web.archive.org/ web/20070920191840/http://www.gammoned.com/snowie.html.

Gates, Bill. 1975. “Software Contest Winners Announced.” *Computer Notes* 1 (2): 1.

Georgieff, Michael K. 2007. “Nutrition and the Developing Brain: Nutrient Priorities and Measurement.” *American Journal of Clinical Nutrition* 85 (2): 614S-620S.

Gianaroli, Luca. 2000. “Preimplantation Genetic Diagnosis: Polar Body and Embryo Biopsy.” Supplement, *Human Reproduction* 15 (4): 69-75.

Gilovich, Thomas, Griffin, Dale, and Kahneman, Daniel, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment.* New York: Cambridge University Press.

Gilster, Paul. 2012. “ESO: Habitable Red Dwarf Planets Abundant.” *Centauri Dreams* (blog), March 29.

Goldstone, Jack A. 1980. “Theories of Revolution: The Third Generation.” *World Politics* 32 (3): 425-53.

Goldstone, Jack A. 2001. “Towards a Fourth Generation of Revolutionary Theory.” *Annual Review of Political Science* 4: 139-87.

Good, Irving John. 1965. “Speculations Concerning the First Ultraintelligent Machine.” In *Advances in Com- puters*, edited by Franz L. Alt and Morris Rubinoff, 6: 31-88. New York: Academic Press.

Good, Irving John. 1970. “Some Future Social Repercussions of Computers.” *International Journal of Environ- mental Studies* 1 (1-4): 67-79.

Good, Irving John. 1976. “Book review of ‘The Thinking Computer: Mind Inside Matter’” In *International Journal of Man-Machine Studies* 8: 617-20.

Good, Irving John. 1982. “Ethical Machines.” In *Intelligent Systems: Practice and Perspective,* edited by J. E. Hayes, Donald Michie, and Y.-H. Pao, 555-60. Machine Intelligence 10. Chichester: Ellis Horwood.

Goodman, Nelson. 1954. *Fact, Fiction, and Forecast.* 1st ed. London: Athlone Press.

Gott, J. R., Juric, M., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N., and Brinkmann, J. 2005.” A Map of the Universe.” *Astrophysical Journal* 624 (2): 463-83.

Gottfredson, Linda S. 2002. “G: Highly General and Highly Practical.” In *The General Factor of Intelligence: How GeneralIs It?,* edited by Robert J. Sternberg and Elena L. Grigorenko, 331-80. Mahwah, NJ: Lawrence Erlbaum.

Gould, S. J. 1990. *Wonderful Life: The Burgess Shale and the Nature of History.* New York: Norton.

Graham, Gordon. 1997. *The Shape of the Past: A Philosophical Approach to History*. New York: Oxford Uni- versity Press.
Gray, C. M., and McCormick, D. A. 1996. "Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex." *Science* 274 (5284): 109-13.
Greene, Kate. 2012. "Intel's Tiny Wi-Fi Chip Could Have a Big Impact." *MIT Technology Review,* September 21.
Guizzo, Erico. 2010. "World Robot Population Reaches 8.6 Million." *IEEE Spectrum,* April 14.
Gunn, James E. 1982. *Isaac Asimov: The Foundations of Science Fiction*. Science-Fiction Writers. New York: Oxford University Press.
Haberl, Helmut, Erb, Karl-Heinz, and Krausmann, Fridolin. 2013. "Global Human Appropriation of Net Pri- mary Production (HANPP)." *Encyclopedia of Earth*, September 3.
Haberl, H., Erb, K. H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W., and Fis- cher-Kowalski, M. 2007. "Quantifying and Mapping the Human Appropriation of Net Primary Produc- tion in Earth's Terrestrial Ecosystems." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12942-7.
Hájek, Alan. 2009. "Dutch Book Arguments." In Anand, Pattanaik, and Puppe 2009, 173-95.
Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine*. Amherst, NY: Prometheus Books.
Hampson, R. E., Song, D., Chan, R. H., Sweatt, A. J., Riley, M. R., Gerhardt, G. A., Shin, D. C., Marmarelis, V. Z., Berger, T. W., and Deadwyler, S. A. 2012. "A Nonlinear Model for Hippocampal Cognitive Prosthesis: Memory Facilitation by Hippocampal Ensemble Stimulation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2): 184-97.
Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2).
Hanson, Robin. 1995. "Could Gambling Save Science? Encouraging an Honest Consensus." *Social Epistemo- logy* 9 (1): 3-33.
Hanson, Robin. 1998a. "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Coloniza- tion." Unpublished manuscript, July 1. Retrieved April 26, 2012. http://hanson.gmu.edu/filluniv.pdf.
Hanson, Robin. 1998b. "Economic Growth Given Machine Intelligence." Unpublished manuscript. Retrieved May 15, 2013. Available at http://hanson.gmu.edu/aigrow.pdf.
Hanson, Robin. 1998c. "Long-Term Growth as a Sequence of Exponential Modes." Unpublished manuscript. Last revised December 2000. Available at http://hanson.gmu.edu/longgrow.pdf.
Hanson, Robin. 1998d. "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions." Unpu- blished manuscript, September 23. Retrieved August 12, 2012. Available at http://hanson.gmu.edu/hard- step.pdf.
Hanson, Robin. 2000. "Shall We Vote on Values, But Bet on Beliefs?" Unpublished manuscript, September. Last revised October 2007. Available at http://hanson.gmu.edu/futarchy.pdf.
Hanson, Robin. 2006. "Uncommon Priors Require Origin Disputes." *Theory and Decision* 61 (4): 319-328.
Hanson, Robin. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45-50.
Hanson, Robin. 2009. "Tiptoe or Dash to Future?" *OvercomingBias* (blog), December 23.
Hanson, Robin. 2012. "Envisioning the Economy, and Society, of Whole Brain Emulations." Paper presented at the AGI Impacts conference 2012.
Hart, Oliver. 2008. "Economica Coase Lecture Reference Points and the Theory of the Firm." *Economica* 75 (299): 404-11.
Hay, Nicholas James. 2005. "Optimal Agents." B.Sc. thesis, University of Auckland.
Hedberg, Sara Reese. 2002. "Dart: Revolutionizing Logistics Planning." *IEEE Intelligent Systems* 17 (3): 81-3.
Helliwell, John, Layard, Richard, and Sachs, Jeffrey. 2012. *World Happiness Report.* The Earth Institute.
Helmstaedter, M., Briggman, K. L., and Denk, W. 2011. "High-Accuracy Neurite Reconstruction for High-Throughput Neuroanatomy." *Nature Neuroscience* 14 (8):

1081-8.
Heyl, Jeremy S. 2005. “The Long-Term Future of Space Travel.” *Physical Review D* 72 (10): 1-4.
Hibbard, Bill. 2011. “Measuring Agent Intelligence via Hierarchies of Environments.” In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Procee- dings,* edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 303-8. Lecture Notes in Computer Science 6830. Berlin: Springer.
Hinke, R. M., Hu, X., Stillman, A. E., Herkle, H., Salmi, R., and Ugurbil, K. 1993. “Functional Magnetic Reso- nance Imaging of Broca’s Area During Internal Speech.” *Neuroreport* 4 (6): 675-8.
Hinxton Group. 2008. *Consensus Statement: Science, Ethics and Policy Challenges of Pluripotent Stem Cell-De- rived Gametes.* Hinxton, Cambridgeshire, UK, April 11. Available at http://www.hinxtongroup.org/Con- sensus_HG08_FINAL.pdf.
Hoffman, David E. 2009. *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy.* New York: Doubleday.
Hofstadter, Douglas R. (1979) 1999. *Godel, Escher, Bach: An Eternal Golden Braid.* New York: Basic Books.
Holley, Rose. 2009. “How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs.” *D-Lib Magazine* 15 (3-4).
Horton, Sue, Alderman, Harold, and Rivera, Juan A. 2008. *Copenhagen Consensus 2008 Challenge Paper: Hunger and Malnutrition.* Technical report. Copenhagen Consensus Center, May 11.
Howson, Colin, and Urbach, Peter. 1993. *Scientific Reasoning: The Bayesian Approach*. 2nd ed. Chicago: Open Court.
Hsu, Stephen. 2012. “Investigating the Genetic Basis for Intelligence and Other Quantitative Traits.” Lecture given at UC Davis Department of Physics Colloquium, Davis, CA, February 13.
Huebner, Bryce. 2008. “Do You See What We See? An Investigation of an Argument Against Collective Representation.” *Philosophical Psychology* 21 (1): 91-112.
Huff, C. D., Xing, J., Rogers, A. R., Witherspoon, D., and Jorde, L. B. 2010. “Mobile Elements Reveal Small Population Size in the Ancient Ancestors of *Homo Sapiens*.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (5): 2147-52.
Huffman, W. Cary, and Pless, Vera. 2003. *Fundamentals of Error-Correcting Codes.* New York: Cambridge University Press.
Hunt, Patrick. 2011. “Late Roman Silk: Smuggling and Espionage in the 6th Century CE.” *Philolog, Stanford University* (blog), August 2.
Hutter, Marcus. 2001. “Towards a Universal Theory of Artificial Intelligence Based on Algorithmic Proba- bility and Sequential Decisions.” In De Raedt and Flach 2001, 226-38.
Hutter, Marcus. 2005. *Universal Artificial Intelligencet: Sequential Decisions Based On Algorithmic Probability.* Texts in Theoretical Computer Science. Berlin: Springer.
Iliadou, A. N., Janson, P. C., and Cnattingius, S. 2011. “Epigenetics and Assisted Reproductive Technology.” *Journal of Internal Medicine* 270 (5): 414-20.
Isaksson, Anders. 2007. *Productivity and Aggregate Growth: A Global Picture.* Technical report 05/2007. Vienna, Austria: UNIDO (United Nations Industrial Development Organization) Research and Statistics Branch.
Jones, Garret. 2009. “Artificial Intelligence and Economic Growth: A Few Finger-Exercises.” Unpublished manuscript, January. Retrieved November 5, 2012. Available at http://mason.gmu.edu/~gjonesb/AIand- Growth.
Jones, Vincent C. 1985. *Manhattan: The Army and the Atomic Bomb*. United States Army in World War II. Washington, DC: Center of Military History.
Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction and Decision Theory. New York: Cambridge University Press.
Judd, K. L., Schmedders, K., and Yeltekin, S. 2012. “Optimal Rules for Patent Races.” *International Economic Review* 53 (1): 23-52.
Kalfoglou, A., Suthers, K., Scott, J., and Hudson, K. 2004. *Reproductive Genetic Testing: What America Thinks.* Genetics and Public Policy Center.
Kamm, Francés M. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm.* Oxford

Ethics Series. New York: Oxford University Press.
Kandel, Eric R., Schwartz, James H., and Jessell, Thomas M., eds. 2000. *Principles of Neural Science.* 4th ed. New York: McGraw-Hill.
Kansa, Eric. 2003. "Social Complexity and Flamboyant Display in Competition: More Thoughts on the Fermi Paradox." Unpublished manuscript, archived version.
Karnofsky, Holden. 2012. "Comment on 'Reply to Holden on Tool AI.' " *Less Wrong* (blog), August 1.
Kasparov, Garry. 1996. "The Day That I Sensed a New Kind of Intelligence." *Time,* March 25, no. 13.
Kaufman, Jeff. 2011. "Whole Brain Emulation and Nematodes." *JeffKaufmans Blog* (blog), November 2.
Keim, G. A., Shazeer, N. M., Littman, M. L., Agarwal, S., Cheves, C. M., Fitzgerald, J., Grosland, J., Jiang, F., Pollard, S., and Weinmeister, K. 1999. "Proverb: The Probabilistic Cruciverbalist." In *Proceedings of the Sixteenth National Conference on Artificial Intelligence,* 710-17. Menlo Park, CA: AAAI Press.
Kell, Harrison J., Lubinski, David, and Benbow, Camilla P. 2013. "Who Rises to the Top? Early Indicators." *Psychological Science* 24 (5): 648-59.
Keller, Wolfgang. 2004. "International Technology Diffusion." *Journal of Economic Literature* 42 (3): 752-82.
KGS Go Server. 2012. "KGS Game Archives: Games of KGS player zen19." Retrieved July 22, 2013. Available at http://www.gokgs.com/gameArchives.jsp?user=zen19d&oldAccounts=t&year=2012&month=3.
Knill, Emanuel, Laflamme, Raymond, and Viola, Lorenza. 2000. "Theory of Quantum Error Correction for General Noise." *Physical Review Letters* 84 (11): 2525-8.
Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V., and Sterling, P. 2006. "How *Much* the Eye Tells the Brain." *Current Biology* 16 (14): 1428-34.
Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., et al. 2012. "Rate of De Novo Mutations and the Importance of Father's Age to Disease Risk." *Nature* 488: 471-5.
Koomey, Jonathan G. 2011. *Growth in Data Center Electricity Use 2005 to 2010.* Technical report, 08/01/2011. Oakland, CA: Analytics Press.
Koubi, Vally. 1999. "Military Technology Races." *International Organization* 53 (3): 537-65.
Koubi, Vally, and Lalman, David. 2007. "Distribution of Power and Military R&D." *Journal of Theoretical Politics* 19 (2): 133-52.
Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., and Lanza, G. 2003. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence.* 2nd ed. Genetic Programming. Norwell, MA: Kluwer Academic.
Kremer, Michael. 1993. "Population Growth and Technological Change: One Million B.C. to 1990." *Quarterly Journal of Economics* 108 (3): 681-716.
Kruel, Alexander. 2011. "Interview Series on Risks from AI." *Less Wrong Wiki* (blog). Retrieved Oct 26, 2013. Available at http://wiki.lesswrong.com/wiki/Interview_series_on_risks_from_AI.
Kruel, Alexander. 2012. "Q&A with Experts on Risks From AI #2." *Less Wrong* (blog), January 9.
Krusienski, D. J., and Shih, J. J. 2011. "Control of a Visual Keyboard Using an Electrocorticographic Brain-Computer Interface." *Neurorehabilitation and NeuralRepair* 25 (4): 323-31.
Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions.* 1st ed. Chicago: University of Chicago Press.
Kuipers, Benjamin. 2012. "An Existing, Ecologically-Successful Genus of Collectively Intelligent Artificial Creatures." Paper presented at the 4th International Conference, ICCCI 2012, Ho Chi Minh City, Vietnam, November 28-30.
Kurzweil, Ray. 2001. "Response to Stephen Hawking." Kurzweil Accelerating Intelligence. September 5. Re- trieved December 31, 2012. Available at http://www.kurzweilai.net/response-to-stephen-hawking.
Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.
Laffont, Jean-Jacques, and Martimort, David. 2002. *The Theory of Incentives: The Principal-*

Agent Model. Princeton, NJ: Princeton University Press.
Lancet, The. 2008. "Iodine Deficiency—Way to Go Yet." *The Lancet* 372 (9633): 88.
Landauer, Thomas K. 1986. "How Much Do People Remember? Some Estimates of the Quantity of Learned Information in Long-Term Memory." *Cognitive Science* 10 (4): 477-93.
Lebedev, Anastasiya. 2004. "The Man Who Saved the World Finally Recognized." *MosNews,* May 21.
Lebedev, M. A., and Nicolelis, M. A. 2006. "Brain-Machine Interfaces: Past, Present and Future." *Trends in Neuroscience* 29 (9): 536-46.
Legg, Shane. 2008. "Machine Super Intelligence." PhD diss., University of Lugano.
Leigh, E. G., Jr. 2010. "The Group Selection Controversy." *Journal of Evolutionary Biology* 23(1): 6-19.
Lenat, Douglas B. 1982. "Learning Program Helps Win National Fleet Wargame Tournament." *SIGART*
Newsletter 79: 16-17.
Lenat, Douglas B. 1983. "EURISKO: A Program that Learns New Heuristics and Domain Concepts." *Artificial Intelligence* 21 (1-2): 61-98.
Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy & Public Affairs* 29 (4): 342-70.
Lerner, Josh. 1997. "An Empirical Exploration of a Technology Race." *RAND Journal of Economics* 28 (2): 228-47.
Leslie, John. 1996. *The End of the World: The Science and Ethics of Human Extinction.* London: Routledge.
Lewis, David. 1988. "Desire as Belief." *Mind: A Quarterly Review of Philosophy* 97 (387): 323-32.
Li, Ming, and Vitányi, Paul M. B. 2008. *An Introduction to Kolmogorov Complexity and Its Applications.* Texts in Computer Science. New York: Springer.
Lin, Thomas, Mausam, and Etzioni, Oren. 2012. "Entity Linking at Web Scale." In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX12),* edited by James Fan, Raphael Hoffman, Aditya Kalyanpur, Sebastian Riedel, Fabian Suchanek, and Partha Pratim Talukdar, 84-88. Madison, WI: Omnipress.
Lloyd, Seth. 2000. "Ultimate Physical Limits to Computation." *Nature* 406 (6799): 1047-54.
Louis Harris & Associates. 1969. "Science, Sex, and Morality Survey, study no. 1927." *Life Magazine* (New York) 4.
Lynch, Michael. 2010. "Rate, Molecular Spectrum, and Consequences of Human Mutation." *Proceedings of the National Academy of Sciences of the United States of America* 107 (3): 961-8.
Lyons, Mark K. 2011. "Deep Brain Stimulation: Current and Future Clinical Applications." *Mayo Clinic Pro- ceedings* 86 (7): 662-72.
MacAskill, William. 2010. "Moral Uncertainty and Intertheoretic Comparisons of Value." BPhil thesis, Uni- versity of Oxford.
McCarthy, John. 2007. "From Here to Human-Level AI." *Artificial Intelligence* 171 (18): 1174-82.
McCorduck, Pamela. 1979. *Machines Who Think: A PersonalInquiry into the History andProspects of Artificial Intelligence*. San Francisco: W. H. Freeman.
Mack, C. A. 2011. "Fifty Years of Moore's Law." *IEEE Transactions on Semiconductor Manufacturing* 24 (2): 202-7.
MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms.* New York: Cambridge University Press.
McLean, George, and Stewart, Brian. 1979. "Norad False Alarm Causes Uproar." *The National*. Aired Novem- ber 10. Ottawa, ON: CBC, 2012. News Broadcast.
Maddison, Angus. 1999. "Economic Progress: The Last Half Century in Historical Perspective." In *Facts and Fancies of Human Development: Annual Symposium and Cunningham Lecture, 1999,* edited by Ian Cas- tles. Occasional Paper Series, 1/2000. Academy of the Social Sciences in Australia.
Maddison, Angus. 2001. *The World Economy: A Millennial Perspective*. Development Centre

Studies. Paris: Development Centre of the Organisation for Economic Co-operation / Development.
Maddison, Angus. 2005. *Growth and Interaction in the World Economy: The Roots of Modernity*. Washington, DC: AEI Press.
Maddison, Angus. 2007. *Contours of the World Economy, 1-2030 AD: Essays in Macro-Economic History*. New York: Oxford University Press.
Maddison, Angus. 2010. "Statistics of World Population, GDP and Per Capita GDP 1-2008 AD." Retrieved Oc- tober 26, 2013. Available at http://www.ggdc.net/maddison/Historical_Statistics/vertical-file_02-2010. xls.
Mai, Q., Yu, Y., Li, T., Wang, L., Chen, M. J., Huang, S. Z., Zhou, C., and Zhou, Q. 2007. "Derivation of Human Embryonic Stem Cell Lines from Parthenogenetic Blastocysts." *Cell Research* 17 (12): 1008-19.
Mak, J. N., and Wolpaw, J. R. 2009. "Clinical Applications of Brain-Computer Interfaces: Current State and Future Prospects." *IEEE Reviews in Biomedical Engineering* 2: 187-99.
Mankiw, N. Gregory. 2009. *Macroeconomics.* 7th ed. New York, NY: Worth.
Mardis, Elaine R. 2011. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470 (7333): 198-203.
Markoff, John. 2011. "Computer Wins on 'Jeopardy!': Trivial, It's Not." *New York Times,* February 16.
Markram, Henry. 2006. "The Blue Brain Project." *Nature Reviews Neuroscience* 7 (2): 153-160.
Mason, Heather. 2003. "Gallup Brain: The Birth of In Vitro Fertilization." *Gallup,* August 5.
Menzel, Randolf, and Giurfa, Martin. 2001. "Cognitive Architecture of a Mini-Brain: The Honeybee." *Trends in Cognitive Sciences* 5 (2): 62-71.
Metzinger, Thomas. 2003. *BeingNo One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.
Mijic, Roko. 2010. "Bootstrapping Safe AGI Goal Systems." Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8.
Mike, Mike. 2013. "Face of Tomorrow." Retrieved June 30, 2012. Available at http://faceoftomorrow.org.
Milgrom, Paul, and Roberts, John. 1990. "Bargaining Costs, Influence Costs, and the Organization of Eco- nomic Activity." In *Perspectives on Positive Political Economy,* edited by James E. Alt and Kenneth A. Shepsle, 57-89. New York: Cambridge University Press.
Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *PsychologicalReview* 63 (2): 81-97.
Miller, Geoffrey. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature.* New York: Doubleday.
Miller, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World.* Dallas, TX: BenBella Books.
Minsky, Marvin. 1967. *Computation: Finite and Infinite Machines.* Englewood Cliffs, NJ: Prentice-Hall.
Minsky, Marvin, ed. 1968. *Semantic Information Processing.* Cambridge, MA: MIT Press.
Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.' " Unpublished manuscript, Octo- ber 1. Retrieved December 31, 2012. Available at http://web.media.mit.edu/~minsky/papers/TrueNames. Afterword.html.
Minsky, Marvin. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* New York: Simon & Schuster.
Minsky, Marvin, and Papert, Seymour. 1969. *Perceptrons: An Introduction to Computational Geometry*. 1st ed. Cambridge, MA: MIT Press.
Moore, Andrew. 2011. "Hedonism." In *The Stanford Encyclopedia of Philosophy,* Winter 2011, edited by Edward N. Zalta. Stanford, CA: Stanford University.
Moravec, Hans P. 1976. "The Role of Raw Power in Intelligence." Unpublished manuscript, May 12. Retrieved August 12, 2012. Available at http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/ l975/Raw.Power.html.
Moravec, Hans P. 1980. "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover." PhD diss., Stanford University.
Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence.* Cambridge,

MA: Harvard University Press.
Moravec, Hans P. 1998. “When Will Computer Hardware Match the Human Brain?” *Journal of Evolution and Technology* 1.
Moravec, Hans P. 1999. “Rise of the Robots.” *Scientific American*, December, 124-35.
Muehlhauser, Luke, and Helm, Louie. 2012. “The Singularity and Machine Ethics.” In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny S0raker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.
Muehlhauser, Luke, and Salamon, Anna. 2012. “Intelligence Explosion: Evidence and Import.” In *Singularity Hypotheses: A Scientific and Philosophical Assessment,* edited by Amnon Eden, Johnny S0raker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.
Müller, Vincent C., and Bostrom, Nick. Forthcoming. “Future Progress in Artificial Intelligence: A Poll Among Experts.” In “Impacts and Risks of Artificial General Intelligence,” edited by Vincent C. Müller, special issue, *Journal of Experimental and Theoretical Artificial Intelligence.*
Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.
Nachman, Michael W., and Crowell, Susan L. 2000. “Estimate of the Mutation Rate per Nucleotide in Hu- mans.” *Genetics* 156 (1): 297-304.
Nagy, Z. P., and Chang, C. C. 2007. “Artificial Gametes.” *Theriogenology* 67 (1): 99-104.
Nagy, Z. P., Kerkis, I., and Chang, C. C. 2008. “Development of Artificial Gametes.” *Reproductive BioMedicine Online* 16 (4): 539-44.
NASA. 2013. “International Space Station: Facts and Figures.” Available at http://www.nasa.gov/worldbook/ intspacestation_worldbook.html.
Newborn, Monty. 2011. *Beyond Deep Blue: Chess in the Stratosphere.* New York: Springer.
Newell, Allen, Shaw, J. C., and Simon, Herbert A. 1958. “Chess-Playing Programs and the Problem of Com- plexity.” *IBM Journal of Research and Development* 2 (4): 320-35.
Newell, Allen, Shaw, J. C., and Simon, Herbert A. 1959. “Report on a General Problem-Solving Program: Pro- ceedings of the International Conference on Information Processing.” In *Information Processing*, 256-64. Paris: UNESCO.
Nicolelis, Miguel A. L., and Lebedev, Mikhail A. 2009. “Principles of Neural Ensemble Physiology Underlying the Operation of Brain-Machine Interfaces.” *Nature Reviews Neuroscience* 10 (7): 530-40.
Nilsson, Nils J. 1984. *Shakey the Robot,* Technical Note 323. Menlo Park, CA: AI Center, SRI International, April.
Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* New York: Cambridge University Press.
Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., and Turkheimer, E. 2012. “Intelli- gence: New Findings and Theoretical Developments.” *American Psychologist* 67 (2): 130-59.
Niven, Larry. 1973. “The Defenseless Dead.” In *Ten Tomorrows,* edited by Roger Elwood, 91-142. New York: Fawcett.
Nordhaus, William D. 2007. “Two Centuries of Productivity Growth in Computing.” *Journal of Economic History* 67 (1): 128-59.
Norton, John D. 2011. “Waiting for Landauer.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42 (3): 184-98.
Olds, James, and Milner, Peter. 1954. “Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain.” *Journal of Comparative and Physiological Psychology* 47 (6): 419-27.
Olum, Ken D. 2002. “The Doomsday Argument and the Number of Possible Observers.” *Philosophical Quar- terly* 52 (207): 164-84.
Omohundro, Stephen M. 2007. “The Nature of Self-Improving Artificial Intelligence.” Paper presented at Sin- gularity Summit 2007, San Francisco, CA, September 8-9.
Omohundro, Stephen M. 2008. “The Basic AI Drives.” In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483-92. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.
Omohundro, Stephen M. 2012. “Rational Artificial Intelligence for the Greater Good.” In

Singularity Hypotheses: A Scientific and Philosophical Assessment, edited by Amnon Eden, Johnny S0raker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

O'Neill, Gerard K. 1974. "The Colonization of Space." *Physics Today* 27 (9): 32-40.

Oshima, Hideki, and Katayama, Yoichi. 2010. "Neuroethics of Deep Brain Stimulation for Mental Disorders: Brain Stimulation Reward in Humans." *Neurologia medico-chirurgica* 50 (9): 845-52.

Parfit, Derek. 1986. *Reasons and Persons.* New York: Oxford University Press.

Parfit, Derek. 2011. *On What Matters.* 2 vols. The Berkeley Tanner Lectures. New York: Oxford University Press.

Parrington, Alan J. 1997. "Mutually Assured Destruction Revisited." *Airpower Journal* 11 (4).

Pasqualotto, Emanuele, Federici, Stefano, and Belardinelli, Marta Olivetti. 2012. "Toward Functioning and Usable Brain-Computer Interfaces (BCIs): A Literature Review." *Disability and Rehabilitation: Assistive Technology* 7 (2): 89-103.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference.* 2nd ed. New York: Cambridge University Press.

Perlmutter, J. S., and Mink, J. W. 2006. "Deep Brain Stimulation." *Annual Review of Neuroscience* 29: 229-57.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined.* New York: Viking.

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control Consortium 2, and Davis, O. S. 2013. "Common DNA Markers Can Account for More than Half of the Genetic Influence on Cognitive Abilities." *Psychological Science* 24 (2): 562-8.

Popper, Nathaniel. 2012. "Flood of Errant Trades Is a Black Eye for Wall Street." *New York Times,* August 1.

Pourret, Olivier, Naim, Patrick, and Marcot, Bruce, eds. 2008. *Bayesian Networks: A Practical Guide to Applications.* Chichester, West Sussex, UK: Wiley.

Powell, A., Shennan, S., and Thomas, M. G. 2009. "Late Pleistocene Demography and the Appearance of Modern Human Behavior." *Science* 324 (5932): 1298-1301.

Price, Huw. 1991. "Agency and Probabilistic Causality." *British Journal for the Philosophy of Science* 42 (2): 157-76.

Qian, M., Wang, D., Watkins, W. E., Gebski, V., Yan, Y. Q., Li, M., and Chen, Z. P. 2005. "The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China." *Asia Pacific Journal of Clinical Nutrition* 14 (1): 32-42.

Quine, Willard Van Orman, and Ullian, Joseph Silbert. 1978. *The Web of Belief,* ed. Richard Malin Ohmann, vol. 2. New York: Random House.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5-31.

Rajab, Moheeb Abu, Zarfoss, Jay, Monrose, Fabian, and Terzis, Andreas. 2006. "A Multifaceted Approach to Understanding the Botnet Phenomenon." In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, 41-52. New York: ACM.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Belknap.

Read, J. I., and Trentham, Neil. 2005. "The Baryonic Mass Function of Galaxies." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363 (1837): 2693-710.

Repantis, D., Schlattmann, P., Laisney, O., and Heuser, I. 2010. "Modafinil and Methylphenidate for Neu- roenhancement in Healthy Individuals: A Systematic Review." *Pharmacological Research* 62 (3): 187-206.

Rhodes, Richard. 1986. *The Making of the Atomic Bomb.* New York: Simon & Schuster.

Rhodes, Richard. 2008. *Arsenals of Folly: The Making of the Nuclear Arms Race*. New York: Vintage.

Rietveld, Cornelius A., Medland, Sarah E., Derringer, Jaime, Yang, Jian, Esko, Tonu, Martin, Nicolas W., Westra, Harm-Jan, Shakhbazov, Konstantin, Abdellaoui, Abdel, et al. 2013. "GWAS of 126,559 Individ- uals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340 (6139): 1467-71.

Ring, Mark, and Orseau, Laurent. 2011. "Delusion, Survival, and Intelligent Agents." In *Artificial General Intelligence: 4th International Conference, AGI2011, Mountain View, CA,*

USA, August 3-6,2011. Proceed- ings, edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 11-20. Lecture Notes in Computer Science 6830. Berlin: Springer.

Ritchie, Graeme, Manurung, Ruli, and Waller, Annalu. 2007. "A Practical Application of Computational Hu- mour." In *Proceedings of the 4th International Joint Workshop on Computational Creativity,* edited by Amilcar Cardoso and Geraint A. Wiggins, 91-8. London: Goldsmiths, University of London.

Roache, Rebecca. 2008. "Ethics, Speculation, and Valúes." *NanoEthics* 2 (3): 317-27.

Robles, J. A., Lineweaver, C. H., Grether, D., Flynn, C., Egan, C. A., Pracy, M. B., Holmberg, J., and Gardner, E. 2008. "A Comprehensive Comparison of the Sun to Other Stars: Searching for Self-Selection Effects." *Astrophysical Journal* 684 (1): 691-706.

Roe, Anne. 1953. *The Making of a Scientist.* New York: Dodd, Mead.

Roy, Deb. 2012. "About." Retrieved October 14. Available at http://web.media.mit.edu/~dkroy/.

Rubin, Jonathan, and Watson, Ian. 2011. "Computer Poker: A Review." *Artificial Intelligence* 175 (5-6): 958-87.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533-6.

Russell, Bertrand. 1986. "The Philosophy of Logical Atomism." In *The Philosophy of Logical Atomism and Other Essays 1914-1919,* edited by John G. Slater, 8: 157-244. The Collected Papers of Bertrand Russell. Boston: Allen & Unwin.

Russell, Bertrand, and Griffin, Nicholas. 2001. *The Selected Letters of Bertrand Russell: The Public Years, 19141970.* New York: Routledge.

Russell, Stuart J., and Norvig, Peter. 2010. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Sabrosky, Curtis W. 1952. "How Many Insects Are There?" In *Insects,* edited by United States Department of Agriculture, 1-7. Yearbook of Agriculture. Washington, DC: United States Government Printing Office.

Salamon, Anna. 2009. "When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth." Working Paper, December 27.

Salem, D. J., and Rowan, A. N. 2001. *The State of the Animals: 2001*. Public Policy Series. Washington, DC: Humane Society Press.

Salverda, W., Nolan, B., and Smeeding, T. M. 2009. *The Oxford Handbook of Economic Inequality.* Oxford: Oxford University Press.

Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210-19.

Sandberg, Anders. 1999. "The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains." *Journal of Evolution and Technology* 5.

Sandberg, Anders. 2010. "An Overview of Models of Technological Singularity." Paper presented at the Road- maps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8.

Sandberg, Anders. 2013. "Feasibility of Whole Brain Emulation." In *Philosophy and Theory of Artificial Intelligence,* edited by Vincent C. Müller, 5: 251-64. Studies in Applied Philosophy, Epistemology and Rational Ethics. New York: Springer.

Sandberg, Anders, and Bostrom, Nick. 2006. "Converging Cognitive Enhancements." *Annals of the New York Academy of Sciences* 1093: 201-27.

Sandberg, Anders, and Bostrom, Nick. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report 2008-3. Future of Humanity Institute, University of Oxford.

Sandberg, Anders, and Bostrom, Nick. 2011. *Machine Intelligence Survey*. Technical Report 2011-1. Future of Humanity Institute, University of Oxford.

Sandberg, Anders, and Savulescu, Julian. 2011. "The Social and Economic Impacts of Cognitive Enhance- ment." In *Enhancing Human Capacities,* edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane, 92-112. Malden, MA: Wiley-Blackwell.

Schaeffer, Jonathan. 1997. *One Jump Ahead: Challenging Human Supremacy in Checkers.* New York: Springer.

Schaeffer, J., Burch, N., Bjornsson, Y., Kishimoto, A., Muller, M., Lake, R., Lu, P., and Sutphen, S. 2007. "Check- ers Is Solved." *Science* 317 (5844): 1518-22.

Schalk, Gerwin. 2008. "Brain-Computer Symbiosis." *Journal of Neural Engineering* 5 (1): P1-P15.

Schelling, Thomas C. 1980. *The Strategy of Conflict.* 2nd ed. Cambridge, MA: Harvard University Press.

Schultz, T. R. 2000. "In Search of Ant Ancestors." *Proceedings of the National Academy of Sciences of the United States of America* 97 (26): 14028-9.

Schultz, W., Dayan, P., and Montague, P. R. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593-9.

Schwartz, Jacob T. 1987. "Limits of Artificial Intelligence." In *Encyclopedia of Artificial Intelligence,* edited by Stuart C. Shapiro and David Eckroth, 1: 488-503. New York: Wiley.

Schwitzgebel, Eric. 2013. "If Materialism is True, the United States is Probably Conscious." Working Paper, February 8.

Sen, Amartya, and Williams, Bernard, eds. 1982. *Utilitarianism and Beyond.* New York: Cambridge Univer- sity Press.

Shanahan, Murray. 2010. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. New York: Oxford University Press.

Shannon, Robert V. 2012. "Advances in Auditory Prostheses." *Current Opinion in Neurology* 25 (1): 61-6.

Shapiro, Stuart C. 1992. "Artificial Intelligence." In *Encyclopedia of Artificial Intelligence,* 2nd ed., 1: 54-7. New York: Wiley.

Sheppard, Brian. 2002. "World-Championship-Caliber Scrabble." *Artificial Intelligence* 134 (1-2): 241-75.

Shoemaker, Sydney. 1969. "Time Without Change." *Journal of Philosophy* 66 (12): 363-81.

Shulman, Carl. 2010a. *Omohundro's "Basic AI Drives" and Catastrophic Risks.* San Francisco, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2010b. *Whole Brain Emulation and the Evolution of Superorganisms.* San Francisco, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2012. "Could We Use Untrustworthy Human Brain Emulations to Make Trustworthy Ones?" Paper presented at the AGI Impacts conference 2012.

Shulman, Carl, and Bostrom, Nick. 2012. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19 (7-8): 103-30.

Shulman, Carl, and Bostrom, Nick. 2014. "Embryo Selection for Cognitive Enhancement: Curiosity or
Game-Changer?" *Global Policy* 5 (1): 85-92.

Shulman, Carl, Jonsson, Henrik, and Tarleton, Nick. 2009. "Which Consequentialism? Machine Ethics and Moral Divergence." In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, Oc- tober 1st-2nd, University of Tokyo, Japan. Proceedings,* edited by Carson Reynolds and Alvaro Cassinelli, 23-25. AP-CAP 2009.

Sidgwick, Henry, and Jones, Emily Elizabeth Constance. 2010. *The Methods of Ethics.* Charleston, SC: Nabu Press.

Silver, Albert. 2006. "How Strong Is GNU Backgammon?" Backgammon Galore! September 16. Retrieved October 26, 2013. Available at http://www.bkgm.com/gnu/AllAboutGNU.html#how_strong_is_gnu.

Simeral, J. D., Kim, S. P., Black, M. J., Donoghue, J. P., and Hochberg, L. R. 2011. "Neural Control of Cursor Trajectory and Click by a Human with Tetraplegia 1000 Days after Implant of an Intracortical Microelec- trode Array." *Journal of Neural Engineering* 8 (2): 025027.

Simester, Duncan, and Knez, Marc. 2002. "Direct and Indirect Bargaining Costs and the Scope of the Firm." *Journal of Business* 75 (2): 283-304.

Simon, Herbert Alexander. 1965. *The Shape of Automation for Men and Management.* New York: Harper & Row.

Sinhababu, Neil. 2009. "The Humean Theory of Motivation Reformulated and Defended." *Philosophical Re- view* 118 (4): 465-500.

Slagle, James R. 1963. "A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calcu- lus." *Journal of the ACM* 10 (4): 507-20.

Smeding, H. M., Speelman, J. D., Koning-Haanstra, M., Schuurman, P. R., Nijssen, P., van Laar, T., and Schmand, B. 2006. "Neuropsychological Effects of Bilateral STN Stimulation in Parkinson Disease: A Controlled Study." *Neurology* 66 (12): 1830-6.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind: A Quarterly Review of Philosophy* 96 (381): 36-61.

Smith, Michael, Lewis, David, and Johnston, Mark. 1989. “Dispositional Theories of Value.” *Proceedings of the Aristotelian Society* 63: 89-174.

Sparrow, Robert. 2013. “In Vitro Eugenics.” *Journal of Medical Ethics.* doi:10.1136/medethics-2012101200. Published online April 4, 2013. Available at http://jme.bmj.com/content/early/ 2013/02/13/ medethics-2012-101200.full.

Stansberry, Matt, and Kudritzki, Julian. 2012. *Uptime Institute 2012 Data Center Industry Survey*. Uptime Institute.

Stapledon, Olaf. 1937. *Star Maker*. London: Methuen.

Steriade, M., Timofeev, I., Durmuller, N., and Grenier, F. 1998. “Dynamic Properties of Corticothalamic Neu- rons and Local Cortical Interneurons Generating Fast Rhythmic (30-40 Hz) Spike Bursts.” *Journal of Neurophysiology* 79 (1): 483-90.

Stewart, P. W., Lonky, E., Reihman, J., Pagano, J., Gump, B. B., and Darvill, T. 2008. “The Relationship Between Prenatal PCB Exposure and Intelligence (IQ) in 9-Year-Old Children.” *Environmental Health Perspectives* 116 (10): 1416-22.

Sun, W., Yu, H., Shen, Y., Banno, Y., Xiang, Z., and Zhang, Z. 2012. “Phylogeny and Evolutionary History of the Silkworm.” *Science China Life Sciences* 55 (6): 483-96.

Sundet, J., Barlaug, D., and Torjussen, T. 2004. “The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Scores of Norwegian Conscripts During Half a Century.” *Intelligence* 32 (4): 349-62.

Sutton, Richard S., and Barto, Andrew G. 1998. *Reinforcement Learning: An Introduction.* Adaptive Compu- tation and Machine Learning. Cambridge, MA: MIT Press.

Talukdar, D., Sudhir, K., and Ainslie, A. 2002. “Investigating New Product Diffusion Across Products and Countries.” *Marketing Science* 21 (1): 97-114.

Teasdale, Thomas W., and Owen, David R. 2008. “Secular Declines in Cognitive Test Scores: A Reversal of the Flynn Effect.” *Intelligence* 36 (2): 121-6.

Tegmark, Max, and Bostrom, Nick. 2005. “Is a Doomsday Catastrophe Likely?” *Nature* 438: 754.

Teitelman, Warren. 1966. “Pilot: A Step Towards Man-Computer Symbiosis.” PhD diss., Massachusetts Insti- tute of Technology.

Temple, Robert K. G. 1986. *The Genius of China: 3000 Years of Science, Discovery, and Invention.* 1st ed. New York: Simon & Schuster.

Tesauro, Gerald. 1995. “Temporal Difference Learning and TD-Gammon.” *Communications of the ACM* 38 (3): 58-68.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton, NJ: Prince- ton University Press.

Tetlock, Philip E., and Belkin, Aaron. 1996. “Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives.” In *Counterfactual Thought Experiments in World Poli- tics: Logical, Methodological, and Psychological Perspectives*, edited by Philip E. Tetlock and Aaron Belkin, 1-38. Princeton, NJ: Princeton University Press.

Thompson, Adrian. 1997. “Artificial Evolution in the Physical World.” In *Evolutionary Robotics: From In- telligent Robots to Artificial Life,* edited by Takashi Gomi, 101-25. Er ’97. Carp, ON: Applied AI Systems.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., et al. 2006. “Stanley: The Robot That Won the DARPA Grand Challenge.” *Journal of Field Robotics* 23 (9): 661-92.

Trachtenberg, J. T., Chen, B. E., Knott, G. W., Feng, G., Sanes, J. R., Welker, E., and Svoboda, K. 2002. “Long- Term In Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex.” *Nature* 420 (6917): 788-94.

Traub, Wesley A. 2012. “Terrestrial, Habitable-Zone Exoplanet Frequency from *Kepler.”Astrophysical Journal* 745 (1): 1-10.

Truman, James W., Taylor, Barbara J., and Awad, Timothy A. 1993. “Formation of the Adult Nervous System.” In *The Development of* Drosophila Melanogaster, edited by Michael Bate and Alfonso Martinez Arias. Plainview, NY: Cold Spring Harbor Laboratory.

Tuomi, Ilkka. 2002. “The Lives and the Death of Moore’s Law.” *First Monday* 7 (11).

Turing, A. M. 1950. “Computing Machinery and Intelligence.” *Mind* 59 (236): 433-60.

Turkheimer, Eric, Haley, Andreana, Waldron, Mary, D’Onofrio, Brian, and Gottesman, Irving I. 2003. “So- cioeconomic Status Modifies Heritability of IQ in Young Children.” *Psychological Science* 14 (6): 623-8.

Uauy, Ricardo, and Dangour, Alan D. 2006. “Nutrition in Brain Development and Aging: Role of Essential Fatty Acids.” Supplement, *Nutrition Reviews* 64 (5): S24-S33.

Ulam, Stanislaw M. 1958. “John von Neumann.” *Bulletin of the American Mathematical Society* 64 (3): 1-49.

Uncertain Future, The. 2012. “Frequently Asked Questions” The Uncertain Future. Retrieved March 25, 2012. Available at http://www.theuncertainfuture.com/faq.html.

U.S. Congress, Office of Technology Assessment. 1995. *U.S.-Russian Cooperation in Space* ISS-618. Washington, DC: U.S. Government Printing Office, April.

Van Zanden, Jan Luiten. 2003. *On Global Economic History: A Personal View on an Agenda for Future Re- search*. International Institute for Social History, July 23.

Vardi, Moshe Y. 2012. “Artificial Intelligence: Past and Future.” *Communications of the ACM* 55 (1): 5.

Vassar, Michael, and Freitas, Robert A., Jr. 2006. “Lifeboat Foundation Nanoshield.” Lifeboat Foundation. Retrieved May 12, 2012. Available at http://lifeboat.com/ex/nanoshield.

Vinge, Vernor. 1993. “The Coming Technological Singularity: How to Survive in the Post-Human Era.” In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace,* 11-22. NASA Conference Publication 10129. NASA Lewis Research Center.

Visscher, P. M., Hill, W. G., and Wray, N. R. 2008. “Heritability in the Genomics Era: Concepts and Miscon- ceptions.” *Nature Reviews Genetics* 9 (4): 255-66.

Vollenweider, Franz, Gamma, Alex, Liechti, Matthias, and Huber, Theo. 1998. “Psychological and Cardiovascular Effects and Short-Term Sequelae of MDMA (‘Ecstasy’) in MDMA-Naive Healthy Volunteers.” *Neuropsychopharmachology* 19 (4): 241-51.

Wade, Michael J. 1976. “Group Selections Among Laboratory Populations of Tribolium.” *Proceedings of the National Academy of Sciences of the United States of America* 73 (12): 4604-7.

Wainwright, Martin J., and Jordan, Michael I. 2008. “Graphical Models, Exponential Families, and Varia- tional Inference.” *Foundations and Trends in Machine Learning* 1 (1-2): 1-305.

Walker, Mark. 2002. “Prolegomena to Any Future Philosophy.” *Journal of Evolution and Technology* 10 (1).

Walsh, Nick Paton. 2001. “Alter our DNA or robots will take over, warns Hawking.” *The Observer,* September 1. http://www.theguardian.com/uk/2001/sep/02/medicalscience.genetics.

Warwick, Kevin. 2002. *I, Cyborg.* London: Century.

Wehner, M., Oliker, L., and Shalf, J. 2008. “Towards Ultra-High Resolution Models of Climate and Weather.” *International Journal of High Performance Computing Applications* 22 (2): 149-65.

Weizenbaum, Joseph. 1966. “Eliza: A Computer Program for the Study of Natural Language Communication Between Man And Machine.” *Communications of the ACM* 9 (1): 36-45.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation.* San Franc- York, CA: W. H. Freeman.

Werbos, Paul John. 1994. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting.* New York: Wiley.

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. 1986. “The Structure of the Nervous System of the Nematode *Caenorhabditis Elegans*.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 314 (1165): 1-340.

Whitehead, Hal. 2003. *Sperm Whales: Social Evolution in the Ocean.* Chicago: University of Chicago Press.

Whitman, William B., Coleman, David C., and Wiebe, William J. 1998. “Prokaryotes: The Unseen Majority.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578-83.

Wiener, Norbert. 1960. “Some Moral and Technical Consequences of Automation.” *Science* 131 (3410): 1355-8.

Wikipedia. 2012a, s.v. “Computer Bridge.” Retrieved June 30, 2013. Available at http://en.wikipedia.org/wiki/ Computer_bridge.

Wikipedia. 2012b, s.v. “Supercomputer.” Retrieved June 30, 2013. Available at http://et.wikipedia.org/wiki/ Superarvuti.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton Science Library. Princeton, NJ: Princeton University

Press.
Winograd, Terry. 1972. *Understanding Natural Language.* New York: Academic Press.

Wood, Nigel. 2007. *Chinese Glazes: Their Origins, Chemistry and Re-creation.* London: A. & C. Black.

World Bank. 2008. *Global Economic Prospects: Technology Diffusion in the Developing World* 42097. Washington, DC.

World Robotics. 2011. *Executive Summary of 1. World Robotics 2011 Industrial Robots; 2. World Robotics 2011 Service Robots.* Retrieved June 30, 2012. Available at http:// www.bara.org.uk/pdf/2012/world-robotics/ Executive_Summary_WR_2012.pdf.

World Values Survey. 2008. *WVS 2005-2008.* Retrieved 29 October, 2013. Available at http:// www.wvsevsdb. com/wvs/WVSAnalize Study.j sp.

Wright, Robert. 2001. *Nonzero: The Logic of Human Destiny.* New York: Vintage.

Yaeger, Larry. 1994. "Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or PolyWorld: Life in a New Context." In *Proceedings of the Artificial Life III Conference,* edited by C. G. Langton, 263-98. Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley.

Yudkowsky, Eliezer. 2001. *CreatingFriendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* Machine Intelligence Research Institute, San Francisco, CA, June 15.

Yudkowsky, Eliezer. 2002. "The A I-Box Experiment." Retrieved January 15, 2012. Available at http:// yudkowsky.net/singularity/aibox.

Yudkowsky, Eliezer. 2004. *CoherentExtrapolated Volition.* Machine Intelligence Research Institute, San Francisco, CA, May.

Yudkowsky, Eliezer. 2007. "Levels of Organization in General Intelligence." In *Artificial General Intelligence,* edited by Ben Goertzel and Cassio Pennachin, 389-501. Cognitive Technologies. Berlin: Springer.

Yudkowsky, Eliezer. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Cirkovic, 308-45. New York: Oxford University Press.

Yudkowsky, Eliezer. 2008b. "Sustained Strong Recursion." *Less Wrong* (blog), December 5.

Yudkowsky, Eliezer. 2010. *Timeless Decision Theory.* Machine Intelligence Research Institute, San Francisco,
CA.

Yudkowsky, Eliezer. 2011. *Complex Value Systems are Required to Realize Valuable Futures*. San Francisco, CA: Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2013. *Intelligence Explosion Microeconomics,* Technical Report 2013-1. Berkeley, CA: Machine Intelligence Research Institute.

Zahavi, Amotz, and Zahavi, Avishag. 1997. *The Handicap Principle: A MissingPiece of Darwirís Puzzle.* Trans- lated by N. Zahavi-Ely and M. P. Ely. New York: Oxford University Press.

Zalasiewicz, J., Williams, M., Smith, A., Barry, T. L., Coe, A. L., Bown, P. R., Brenchley, P., et al. 2008. "Are We Now Living in the Anthropocene?" *GSA Today* 18 (2): 4-8.

Zeira, Joseph. 2011. "Innovations, Patent Races and Endogenous Growth." *Journal of Economic Growth* 16 (2): 135-56.

Zuleta, Hernando. 2008. "An Empirical Note on Factor Shares." *Journal of International Trade and Economic Development* 17 (3): 379-90.

ÍNDICE

A

acelerador de desarrollo ma- cro-estructural 235, 236 acelerador universal 234
Acumulación de valores 189-190, 208
adquisición de recursos 113114, 123, 153, 193 Afgano Talibán 215 Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA) 16 Agente artificial 88, 105, 107, 166, 172, 185
Agente bayesiano 9-10, 123 ajedrez 12-14, 20, 22, 52, 93, 134, 265-266
alcance directo 58, 92, 280 alcance indirecto 57-58, 92, 99-100, 103, 281 algoritmo de retropropaga- ción 8
algoritmos genéticos 7-9, 13, 24, 27, 238, 241 Andamiaje motivacional 191-192
anticipación de consecuencias (argumentos de) 240 antrópica 127, 134, 143, 157158, 222, 225, 285 aprendizaje artificial 8-9, 1113, 15-16, 18, 28, 188
Aprendizaje de valores 192, 208
aprendizaje por refuerzo 12, 28, 121, 188-189, 194, 207, 238, 279, 284, 292
apuesta Pascaliana 223
Arendt, Hannah 105 Armstrong, Stuart 263, 265267, 280, 282, 286, 293, 296, 303
Asimov, Isaac 139, 286 atraco Pascaliano 223, 300 atrofia 129, 135-137, 143
aumentación 138, 142-143, 201, 203 autista 57, 149
automatización 18, 34, 99, 227-118, 161
auto-mejoramiento recursivo 29, 96, 142

B

backgammon 12, 188 Berliner, Hans 12, 265 Brown, Louise 43, 302
búsqueda exhaustiva 6 búsqueda heurística 6

C

C. elegans 34-35, 268 cables trampa 129, 136-138, 143
capacidad de memoria 7, 72 capital 39, 48, 68, 84, 88, 99, 161-163, 165-167, 170
carga mutacional 41 carrera tecnológica 80-81, 205, 232-233, 248 Carta blanca (juego) 13 Catástrofe de la hipótesis de Riemann 123, 141 Chalmers, David 267, 282, 285, 303, 310-311
CHINOOK 12 Christiano, Paul 200, 208, 295
cláusula de bonanza 255, 305 clonación 41 definición 41 Código alemán Enigma 87
cognición biológica 22, 36, 44, 51, 233
colaboración (beneficios de) 247, 250
comercio algorítmico 18 compartir memoria 61 computronium; *véase también* potencia computa- cional 102-103, 123-124, 193, 219
conexionismo 8 consciencia; *véase también* crimen mental 125, 153, 116, 273
Conjetura de la compleción tecnológica 230 Contenido de los objetivos 222
control de la capacidad 127, 129, 135, 143, 158, 185 Copérnico, Nicolás 14 costes

de negociación 182 crecimiento de 1, 27, 49, 74, 79, 92, 164-165, 263, 275, 283 crecimiento económico 163164, 233, 263, 276, 301 crecimiento moral 214 Crimen mental 125 criptografía 82 crisis ontológica 146 crucigramas (resolver) 12-13 cyborg 44-48, 67, 272

D

damas 12
DARPA, *véase* Agencia de Proyectos de investigación Avanzada de Defensa DART (herramienta) 16 Declaración de Helsinki 188 Deep Blue 12, 319 Deep Fritz 22 degradación elegante 8 demostrador de teoremas 15, 268
Desarrollo tecnológico diferencial (Principio de) 230-231, 238
desempleo 65, 160, 161, 288 Dewey, Daniel 312 dinámica de carrera, *véase* carrera tecnológica diseño institucional 203-204, 206-208
domesticidad 138, 141, 146, 150, 191, 207, 222 Drexler, Eric 239, 272, 278280, 302
drones 15, 99, 117 Dyson, Freeman 312

E

ECC, *véase* emulación de cerebro completo (ECC) economía mundial 2-3, 63, 74, 83, 163-164, 275-276, 279, 287
emulación de cerebro completo (ECC) 28, 30-35, 50, 53, 60, 68-69, 77, 83-84, 108, 172, 179-202, 237239, 241, 245, 253, 268269, 272, 276, 301-302 encajamiento 129-130, 143, 146, 148, 156-157 informacional 130 físico 129-130 enfoque Hail Mary 295 entorno de adaptabilidad evolutiva 164, 171 epistemología 11, 94, 209, 221, 224-225, 227, 259, 294, 299, 300
escaneado, *véase* emulación de cerebro completo (ECC)
escenarios multipolares 131, 159, 180
esfuerzo de diseño, *véase* potencia de optimización especificación directa 138142, 287
esquema de cifrado RSA 82 estatus moral 125-126, 167, 173, 202, 205, 270, 290, 297
eugenesia 36-37, 42, 270 Eurisko 12, 265 evolución 8- 9, 23-28, 51, 100, 104, 155, 159, 164165, 174-175, 187-188, 190, 202, 207-208, 230, 232, 238, 253, 267-269, 275, 280, 288
excedente de hardware 73, 241-243, 291, 303-304 explosión combinatoria 6, 9, 11, 47, 155
explosión de inteligencia 2, 4-5, 29, 62, 64, 75, 77-78, 84-86, 96, 108, 115, 127128, 136, 151, 162, 165, 198, 206-227, 232-233, 236-237, 241-242, 244, 246-247, 250, 252-253, 255, 257, 260, 275-276, 283, 286, 288, 301, 303, 305
extinción humana, *véase* riesgo existencial

F

fallo maligno 120, 123, 149, 196
fertilización in vitro, *véase* selección de embriones fichas de recompensa criptográfica 133
FIV, *véase* selección de embriones
Flash Crash 17-18 función de aptitud 8, 25, 267, *véase también* evolución función de utilidad 10, 88, 110, 119, 124-125, 133134, 172, 185-187, 195196, 200, 208, 292-295, 304-305

G

genio IA 148-158, 279, 286287
genotipado 37 Ginsberg, Matt 13 Giro traicionero 116, 118, 128-129 Go (juego) 13
Good, I.J. 4, 263-265, 267 Gorbachov, Mikhail 86-87

H

Hanson, Robin 160, 263, 272-273, 276, 288, 290,
300, 303
hardware evolucionable 154 Haz lo que quiero que hagas (DWIM - Do What I Mean)
299 hedonismo 140, 210 hedonium 140, 219 Hill, Benny 105 hilos teleológicos 110
hipocampo 47 hipótesis de simulación 134135, 290, 294

I

IA completa (problema de) 145, 186, 249
IA en sentido fuerte 62, 78, 96
IA-herramienta 151, 157158
definición 151 IA-NH, *véase* inteligencia artificial, nivel humano IA-OUM, *véase*
nociones de optimización
IA-RL, *véase* nociones de optimización
IA-VL, *véase* nociones de optimización
IA neuromórfica 33, 47, 50, 238-239, 243, 245, 269,
301, 303
IA productora de clips 123 IA seminal 29, 35, 75, 83, 92, 95, 107, 117-120, 142,
151, 186, 189, 191-192, 197199, 201, 208-211, 214,
217, 224-225, 241, 268, 276-277, 284
IAs dedicadas a juegos 12, 15, 265
implante cerebral, *véase* ci- borg 47-48 Inteligencia artificial
carreras armamentísticas 80, 180, 283, 287 futuro de 18-19, 70, 294 historia de
1, 5 mayor que la humana, *véase* superingeligencia pioneros 4-5, 18
sobrepredicción de 23, 244
Inteligencia artificial a la antigua usanza (GOFAI - Good Old Fashioned Artificial
Intelligence) 7-9, 11, 15
inteligencia colectiva 48-49, 51, 54-56, 67, 72, 142, 164, 203, 273, 275, 280, 304
inteligencia humana 4-5, 11, 14, 18, 24-27, 36, 46, 5758, 63, 74, 78, 98, 124, 160,
172, 243, 257, 271 interfaces de cerebro-ordenador; *véase también* cyborg 44, 46, 48,
51, 67, 83, 142 interfaz hombre-máquina;
véase cyborg 22 internet 16, 19, 45, 49, 71, 73, 77, 85, 94, 96, 98, 130, 137,
138, 145-146, 242, 273 interruptor de parada automático 137-138 intervenciones en
la línea germinal; *véase también* selección de embriones 42, 44, 67, 275
introducción de valores 185, 187-193, 195, 200-201,
207-208, 294-295

J

jaula de Faraday 130 Jeopardy! 13, 71

K

Kasparov, Garry 12, 265 Kepler, Johannes 14 Knuth, Donald 14, 265
Kurzweil, Ray 2, 263, 271

L

La síntesis de ADN 39, 280 Lenat, Douglas 12, 265 lenguaje formal 7, 145 lenguaje natural 14, 71, 140, 145, 218
ley de Moore; *véase también* potencia computacional 24, 27, 73, 76-77, 276, 288
Libros holandeses 111 línea base humana 75, 77, 82 línea de base de la civilización 63 Logistello 12

M

máquina infantil; *véase también* IA seminal 23, 29 McCarthy, John 5, 13, 16, 18, 266, 279
Medalla Fields 256, 273 MegaTierra 55-56 mejora biológica 44, 50, 142, 260
mejora cognitiva 44, 47, 51, 66-67, 94, 111-112, 193, 233-237, 239, 270, 283 metas autolimitantes 123 Método de Monte Carlo 11, 13
métodos de control; *véase también* el control de la capacidad y la selección de la motivación 129-133, 135-136, 143-144, 157159, 191, 206, 208, 257, 285, 287
Métodos de incentivos 129, 131, 133, 137, 143 fichas de recompensa criptográfica 133, 135, 285
integración social 131132, 143, 157-159, 203, 285
Mill, John Stuart 210
minería de datos 233, 302 Minsky, Marvin 18, 263-264, 266, 284
modelado neurocomputacio- nal; *véase también* emulación de cerebro completo (ECC) e IA neuromórfica 35, 302
Modelo de Hodgkin-Huxley 25
modelos gráficos 9, 11 modos de fallo 120, 126, 149 Modulación de emulaciones 201, 208
Moravec, Hans 264, 267, 290

N

Naciones Unidas 87-89, 253254
nanotecnología 94, 97-98, 100, 103, 113, 177, 232, 239, 278-279, 301-302 neurona de McCulloch-Pitts 238
Newton, Isaac 56 Nilsson, Nils 18-19, 264-266 nociones de optimización 11 Agente bayesiano 9-10, 123
aprendiz de valores (IA- VL) 195
aprendiz por refuerzo (IA-RL) 194
maximizador de la utilidad observacional (IA- OUM) 291 nootrópicos 36, 67
normatividad indirecta 138, 141-143, 150, 209-212, 217, 221-222, 264, 300 Norvig, Peter 19, 265-266, 282, 284

O

objetivo basado en razones 220
Oliphant, Mark 85 O'Neill, Gerard 102, 280 oráculo IA 141, 145-150, 152-153, 156, 158, 222, 226, 287-288

P

paradigma logicista, *véase* inteligencia artificial a la antigua usanza Parfit, Derek 281
permisibilidad moral (PM) 218-219, 221, 299 perspectiva de la persona afectada 229, 246-247,

303
perspectiva impersonal 229, 232, 241, 247
pinchazo cerebral 122-123, 133, 189, 194-195, 207, 284, 292
plasticidad del cerebro 48 poker 13
potencia computacional; *véase también* computronium y excedente de hardware 24, 73, 77, 102, 200 potencia de optimización 62, 74-77, 92 definición 74
predecir el comportamiento 121, 206
predicción errónea 4 Principio de deferencia epis- témica 211, 221 Principio del bien común 255, 260
procesamiento sub-simbóli- co, *véase* conexionismo 8 problema de la primacía de agencia 127-128 profusión infraestructural 123-125, 153, 187, 226, 284
programación 29 Protocolo de intercambio de claves Diffie Hellman 82 Proverb (programa) 12 Proyecto Dartmouth Sum- mer 5
Proyecto de sistemas informáticos de quinta generación 7
Proyecto genoma humano
86, 254, 278
Proyecto Manhattan 84-85,
87, 278
punto de Schelling 147, 184, 286, 297

Q

qualia, *véase* consciencia

R

ratificación 157, 225-227 Rawls, John 150, 287, 297, 305
Reagan, Ronald 86-87 realidad virtual 30, 53, 113, 166, 171, 200, 205, 302
reconocimiento de voz 1516, 46
reconocimiento facial 15 Redes bayesianas 9 redes neuronales 5, 7-9, 28, 46, 137, 173, 238, 264, 275-276
reconocimiento de caracteres 15
reconocimiento de voz 1516, 46
reconocimiento facial 15, 292 rectitud moral (RM) 217218, 298-299
Recursos cósmicos 101-102, 104, 112, 114-115, 134, 209, 214, 217, 227, 251, 261, 284, 297
rendimientos decrecientes 37, 66, 88, 275, 304 representación explícita 172
resistencia al progreso 65-77, 92, 241, 276
Revolución Agrícola 2, 80, 263
Revolución Industrial 2, 161, 164
riesgo existencial 86, 103, 115, 152, 231, 234, 241, 242, 244-245, 247, 257, 260-261, 303-304 riesgos de estado 234-235 riesgos de transición 234235, 302
ritmo de crecimiento, *véase* crecimiento 1-2, 74, 164, 263
robótica 11, 19, 117-118, 139, 238, 278 Roosevelt, Franklin D. 85 Russell, Bertrand 6, 87, 139, 265, 278-279, 282, 284, 286

S

salarios 65, 160-162, 167, 289
Samuel, Arthur 12, 265 Sandberg, Anders 307, 309, 311, 321
Schaeffer, Jonathan 321 Scrabble 13, 322 segunda transición 177, 239, 244, 246, 253
semánticas de validez causales 198
Shakey (robot) 6 SHRDLU (programa) 6 Shulman, Carl 267-269 selección de embriones 3741, 43, 50, 67, 269-270 Selección de motivación 131132, 138, 143,

146, 150, 158, 168, 182, 185, 187, 285, 296
Selección evolutiva 187, 207 selección genética; *véase también* evolución 37, 39-40, 42-43, 50, 61, 233, 239 semántica de referencia externa 210
sentido común 3, 14, 55, 117, 139, 260, 266
señal de recompensa 121122, 188, 194, 207 señalización social 110-112, 171, 292
sesgo inductivo 10 Shulman, Carl 269-270, 282, 285, 289, 291-292, 296, 301, 303-305
singularidad; *véase también* explosión de inteligencia 1-2, 49, 76, 263, 276 sistemas de apoyo a la decisión; *véase* también IA-herramienta sistemas de asistencia 15
sistemas de control de inventario 16
sistema experto 7 Situación maltusiana 163, 165 soberano IA 226, 287
solucionadores de ecuaciones 15
solucionadores de problemas 16
sonda von Neumann 283 sopa algorítmica 173 subida a la nube; *véase* emulación de cerebro completo (ECC) 30, 167 Superinteligencia; *véase también* superinteligencia colectiva, superinteligencia de calidad y superinteli- gencia de velocidad definición 22 formas 22, 52, 58, 272-273 caminos hacia 22, 267 predecir el comportamiento de
superinteligencia colectiva 39, 48-49, 51-56, 58, 83, 93, 99, 287 definición 53
Superinteligencia de calidad 52, 56-58, 243 definición 56
Superinteligencia de velocidad 52-54, 58, 76, 272273
superorganismos 178-180 superpoderes 91, 93, 96, 100, 104, 133, 279
    tipos 91, 93, 96, 100, 104, 133, 279
Suplantación perversa 120124, 153, 197 Szilárd, Leó 85

T

TD-Gammon 12 teoría de autómatas 5 teoría de juegos 87, 159, 184
teoría de la decisión; *ver también* nociones de optimización 209, 221-225, 227, 282, 299-300
teoría de la selección obser- vacional, *véase* antrópicos teórico lógico (sistema) 6
terapia génica somática 42, 275
Tesauro, Gerry 12, 265 tesis de la convergencia instrumental 108-109, 115116
tesis de ortogonalidad 105, 107-109, 115, 281 TextRunner (sistema) 71 tres leyes de la robótica 139 Thrun, Sebastián 19, 264 Traducción automática 15 Tribolium castaneum 155 Truman, Harry 85, 267 Turing, Alan 4, 23, 28-29, 44, 225, 264, 267, 273

U

umbral de Unidad-sabia de sostenibilidad 104 Unidad 78, 83, 87-91, 100101, 103-104, 113-115,
    119, 159-160, 177-178,
    180, 182, 184, 231, 242, 245, 253, 277-278, 280, 281-283, 289, 301-302, 304-305

V

VCE, *véase* voluntad coherente extrapolada
vehículo no tripulado, *véase* dron
velo de ignorancia 150, 157, 287
ventaja estratégica decisiva 78-79, 82-83, 87-89, 95, 104, 112, 115-117, 119126,

129, 131, 135, 138,
148, 156-157, 159, 177178, 190, 209, 214, 225,
253
vigilancia 15, 49, 64, 79, 8485, 94, 117, 132, 181, 184, 233, 254, 278, 296, 301302
Vinge, Vernor 2, 49, 263, 272 visión por ordenador 34 voluntad coherente
extrapolada (VCE) 211-222, 226227, 297-300
von Neumann, John 44, 87, 100, 101, 113-114, 263, 279, 283

W

Watson (IBM) 13, 71, 265 Whitehead, Alfred N. 6, 274 Wigner, Eugene 85 Winston,
Patrick 18, 266

Y

Yudkowsky, Eliezer 70, 92, 106, 197-198, 211-214,
216, 264, 268, 272, 275276, 280-281, 284-285, 288-290, 292-293, 295, 297,
300